USA HUMAN
GENOME PROJEC

THE HUMAN GENOME PROJECT: HOW PRIVATE SECTOR DEVELOPMENTS AFFECT THE GOVERNMENT PROGRAM

HEARING

BEFORE THE

SUBCOMMITTEE ON ENERGY AND ENVIRONMENT

OF THE

COMMITTEE ON SCIENCE

U.S. HOUSE OF REPRESENTATIVES

ONE HUNDRED FIFTH CONGRESS

SECOND SESSION

JUNE 17, 1998

[No. 66]

Printed for the use of the Committee on Science

Sanger Centre
Library

BOOKS EXPRESS

Official distributors of books and documents for United States, Australian, Canadian governments and international organisations

P.O. Box 10, Saffron Walden,
Essex CB11 4EW. U.K.
Tel: 01799 513726
Fax: 01799 513248
E-mail: info@books-express.co.uk
URL: www.books-express.co.uk

22502807834

THE HUMAN GENOME PROJECT: HOW PRIVATE SECTOR DEVELOPMENTS AFFECT THE GOVERNMENT PROGRAM

HEARING

BEFORE THE

SUBCOMMITTEE ON ENERGY AND ENVIRONMENT

OF THE

COMMITTEE ON SCIENCE

U.S. HOUSE OF REPRESENTATIVES

ONE HUNDRED FIFTH CONGRESS

SECOND SESSION

JUNE 17, 1998

[No. 66]

Printed for the use of the Committee on Science

The Sanger Centre
54036000014802

U.S. GOVERNMENT PRINTING OFFICE
51–217CC WASHINGTON : 1998

For sale by the U.S. Government Printing Office
Superintendent of Documents, Congressional Sales Office, Washington, DC 20402
ISBN 0-16-057661-X

COMMITTEE ON SCIENCE

F. JAMES SENSENBRENNER, Jr., Wisconsin, *Chairman*

SHERWOOD L. BOEHLERT, New York
HARRIS W. FAWELL, Illinois
CONSTANCE A. MORELLA, Maryland
CURT WELDON, Pennsylvania
DANA ROHRABACHER, California
STEVEN SCHIFF, New Mexico
JOE BARTON, Texas
KEN CALVERT, California
ROSCOE G. BARTLETT, Maryland
VERNON J. EHLERS, Michigan
DAVE WELDON, Florida
MATT SALMON, Arizona
THOMAS M. DAVIS, Virginia
GIL GUTKNECHT, Minnesota
MARK FOLEY, Florida
THOMAS W. EWING, Illinois
CHARLES W. "CHIP" PICKERING, Mississippi
CHRIS CANNON, Utah
KEVIN BRADY, Texas
MERRILL COOK, Utah
PHIL ENGLISH, Pennsylvania
GEORGE R. NETHERCUTT, JR., Washington
TOM A. COBURN, Oklahoma
PETE SESSIONS, Texas

GEORGE E. BROWN, Jr., California RMM*
RALPH M. HALL, Texas
BART GORDON, Tennessee
JAMES A. TRAFICANT, Jr., Ohio
TIM ROEMER, Indiana
ROBERT E. "BUD" CRAMER, Jr., Alabama
JAMES A. BARCIA, Michigan
PAUL MCHALE, Pennsylvania
EDDIE BERNICE JOHNSON, Texas
ALCEE L. HASTINGS, Florida
LYNN N. RIVERS, Michigan
ZOE LOFGREN, California
LLOYD DOGGETT, Texas
MICHAEL F. DOYLE, Pennsylvania
SHEILA JACKSON LEE, Texas
BILL LUTHER, Minnesota
WALTER H. CAPPS, California
DEBBIE STABENOW, Michigan
BOB ETHERIDGE, North Carolina
NICK LAMPSON, Texas
DARLENE HOOLEY, Oregon

Todd R. Schultz, *Chief of Staff*
Barry C. Beringer, *Chief Counsel*
Patricia S. Schwartz, *Chief Clerk/Administrator*
Vivian A. Tessieri, *Legislative Clerk*
Robert E. Palmer, *Democratic Staff Director*

Subcommittee on Energy and Environment

KEN CALVERT, California, *Chairman*

HARRIS W. FAWELL, Illinois
CURT WELDON, Pennsylvania
DANA ROHRABACHER, California
STEVEN H. SCHIFF, New Mexico
VERNON J. EHLERS, Michigan
MATT SALMON, Arizona
MARK ADAM FOLEY, Florida
PHIL ENGLISH, Pennsylvania
TOM A. COBURN, Oklahoma

TIM ROEMER, Indiana
PAUL McHALE, Pennsylvania
MICHAEL F. DOYLE, Pennsylvania
DARLENE HOOLEY, Oregon
RALPH M. HALL, Texas
EDDIE BERNICE JOHNSON, Texas
ZOE LOFGREN, California
ALCEE L. HASTINGS, Florida

---

*Ranking Minority Member
**Vice Chairman

WELLCOME
LIBRARY
General Collectio
M
10188

CONTENTS

Page

June 17, 1998—The Human Genome Project: How Private Sector Developments Affect the Government Program

Opening Statement by Representative Ken Calvert (CA–43), Chairman, Subcommittee on Energy and Environment, Committee on Science, U.S. House of Representatives 1

Opening Statement by Representative Tim Roemer (IN–3), Ranking Minority Member, Subcommittee on Energy and Environment, Committee on Science, U.S. House of Representatives 2

Panel:

Dr. Aristides A. Patrinos, Associate Director of Energy Research for Biological and Environmental Research, U.S. Department of Energy, Washington, DC:
- Oral Testimony 5
- Prepared Testimony 8
- Biography 14

Dr. Francis S. Collins, Director, National Human Genome Research Institute, National Institutes of Health, U.S. Department of Health and Human Services, Bethesda, MD:
- Oral Testimony 15
- Prepared Testimony 18
- Biography 25

Dr. J. Craig Venter, President and Director, The Institute for Genomic Research, Rockville, MD:
- Oral Testimony 26
- Prepared Testimony 28
- Biography 36
- Financial Disclosure 37

Dr. David J. Galas, President and Chief Scientific Officer, Chiroscience R&D Inc., Bothell, WA:
- Oral Testimony 42
- Prepared Testimony 46
- Biography 53
- Financial Disclosure 54

Dr. Maynard V. Olson, Professor of Medical Genetics and Genetics, Department of Molecular Biotechnology, and Director, Genome Center, University of Washington, Seattle, WA:
- Oral Testimony 55
- Prepared Testimony 58
- Biography 64
- Financial Disclosure 71

Discussion
- Reasons for Federal Government To Complete Human Genome Sequencing 72
- Refocusing of Federal Human Genome Project 73
- Federal Program's Use of Latest Technologies 74
- Federal Budget for the Human Genome Project 74
- Dr. Olson's Criticisms of Private-Sector Venture 75
- Ethical, Legal and Social Concerns 77
- Patentability of Human Genome 77
- Difference Between Federal Human Genome Project and Private-Sector Venture 78

Page
Recapturing Private Investment 79
Tension Between Free Market and Information Dissemination 80
Concerns About Public Access to Information 81
Consequences of Intellectual Property/Patient/Privacy Rights 83
Consequences of Private-Sector Venture for Federal Human Genome Project 84
Efficiency of Federal Human Genome Project 84

Appendix 1: Answers to Post-Hearing Questions Submitted by Members of the Subcommittee on Energy and Environment

Dr. Aristides A. Patrinos, Associate Director of Energy Research for Biological and Environmental Research, U.S. Department of Energy:

Republican Member Questions:

Scientific Justification for Completing Government-Funded Sequencing of Entire Human Genome 87
Efficiencies of DOE's Joint Genome Initiative vs. Three Different DOE Laboratory Programs 88

Democratic Member Questions:

Difference Between the DOE–NIH and "Shotgun" Human DNA Sequencing Approaches 89
Role of DOE and NIH in Collaboration with Private-Sector Venture 89
Concerns of International Collaborators About Intellectual Property Rights and Patenting 90

Dr. Francis S. Collins, Director, National Human Genome Research Institute, National Institutes of Health, U.S. Department of Health and Human Services:

Republican Member Question:

Scientific Justification for Completing Government-Funded Sequencing of Entire Human Genome 92

Democratic Member Questions:

Difference Between the DOE–NIH and "Shotgun" Human DNA Sequencing Approaches 93
Role of DOE and NIH in Collaboration with Private-Sector Venture 94
Concerns of International Collaborators About Intellectual Property Rights and Patenting 94
Federal Government's Cost to Completely Sequence the Human Genome 96

Dr. J. Craig Venter, President and Director, The Institute for Genomic Research:

Republican Member Questions:

Will the Private Initiative Duplicate the Federal Human Genome Project? 97
Concern About Release of Data to the Public 98
Recommendations for Restructuring the Federal Human Genome Project 98

Democratic Member Questions:

Availability of Genomic Information to the Scientific Community 99
Timeliness of Release of and Compensation for Human DNA Sequence Data 99
Plans to Patent Genomic Sequences 100
Uniqueness of Expressed Sequence Tags 100
Role of DOE and NIH in Collaboration with Private-Sector Venture 101
Restrictions on Researchers' Ability to Obtain Human DNA Sequence Information 101

Page

Relation of New Venture to the Federally-Funded Human Genome Sequencing Effort 102

Dr. David J. Galas, President and Chief Scientific Officer, Chiroscience R&D Inc:

Republican Member Questions:

Practical Value of Federal Completion of Entire Human Genome Sequencing Process 103

Democratic Member Questions:

Impact on Current Efforts 104

Importance of Genomic Data That May Be Withheld 104

Reasonable Fees and Conditions to Private-Controlled Genetic Information 104

Rights of Individuals' Privacy and Compensation Issues 105

Dr. Maynard V. Olson, Professor of Medical Genetics and Genetics, Department of Molecular Biotechnology, and Director, Genome Center, University of Washington:

Democratic Member Questions:

Concerns About Ability to Access Genomic Information 106

Impact on Current Efforts 107

Importance of Genomic Data That May Be Withheld 107

Reasonable Fees and Conditions to Private-Controlled Genetic Information 107

Rights of Individuals' Privacy and Compensation Issues 108

Appendix 2: Additional Materials for the Record

J. Craig Venter, *et al.*, "Shotgun Sequencing of the Human Genome," *Science* **280,** 1540 (June 5, 1998) 110

Nicholas Wade, "Scientist's Plan: Map All DNA Within 3 Years," *The New York Times,* May 10, 1998, p. A1 113

Bill Richards, "Perkin-Elmer Jumps Into Race to Decode Genes," *The Wall Street Journal,* May 11, 1998, p. B6 115

Nicholas Wade, "Beyond Sequencing of Human DNA," *The New York Times,* May 12, 1998, p. C3 116

Justin Gillis and Rick Weiss, "Private Firm Aims to Beat Government in Gene Plan," *The Washington Post,* May 12, 1998, p. A1 118

Clive Cookson, "Genetic mapping triggers contest: Academics race private enterprise," *The New York Times,* May 12, 1998, p. C16 120

Nicholas Wade, "International Gene Project Gets Lift: Wellcome Trust Doubles Commitment to Public-Sector Effort," *The New York Times,* May 12, 1998, p. A20 121

William A. Haseltine, "Gene-Mapping, Without Tax Money," *The New York Times,* May 21, 1998, p. A37 123

John Carey, "The Duo Jolting the Gene Business: Craig Venter and Perkin-Elmer target the human genome," *Business Week,* May 25, 1998, pp. 70–71 124

Steven E. Koonin, "An Independent Perspective on the Human Genome Project," *Science* **279,** 36 (January 2, 1998) 126

Human Genome Program Report, Part 1, Overview and Progress, Prepared by the Human Genome Management Information System, Oak Ridge National Laboratory for the U.S. Department of Energy, Office of Energy Research, Office of Biological and Environmental Research, DOE/ER–0713 (Part 1), November 1997 128

Human Genome Program Report, Part 2, 1996 Research Abstracts, Prepared by the Human Genome Management Information System, Oak Ridge National Laboratory for the U.S. Department of Energy, Office of Energy Research, Office of Biological and Environmental Research, DOE/ER–0713 (Part 2), November 1997 240

Page

William A. Haseltine, "Discovering Genes for New Medicines," *Scientific American* **276,** No. 3, March 1997, pp. 2–7 338

To Know Ourselves: The U.S. Department of Energy and the Human Genome Project, Prepared by the Lawrence Berkeley National Laboratory for the U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, July 1996 345

Francis Collins and David Galas, "A New Five-Year Plan for the U.S. Human Genome Program," *Science* **262,** 43 (1993) 380

DOE Human Genome Program Primer on Molecular Genetics, Prepared by the Human Genome Management Information System, Oak Ridge National Laboratory for the U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, June 1992 390

THE HUMAN GENOME PROJECT: HOW PRIVATE SECTOR DEVELOPMENTS AFFECT THE GOVERNMENT PROGRAM

WEDNESDAY, JUNE 17, 1998

HOUSE OF REPRESENTATIVES,
COMMITTEE ON SCIENCE,
SUBCOMMITTEE ON ENERGY AND ENVIRONMENT,
Washington, DC.

The Subcommittee met, pursuant to notice, at 1:05 p.m., in room 2318, Rayburn House Office Building, Hon. Ken Calvert, Chairman of the Subcommittee, presiding.

Chairman CALVERT. This hearing of the Energy and Environment Subcommittee will come to order.

Today we will review a program whose success will have profound importance for medical science for the 21st Century. Some of our witnesses today have used some strong language in describing the value of the human genome project, but it's hard to exaggerate the importance of a program that could lead to prevention, and even cures, to some of the most serious diseases that afflict us. The sequencing of the human genome began in the mid-1980's as an effort by the Department of Energy (DOE) to study the effects of radiation on the survivors of Hiroshima and Nagasaki. However, it became an international program with much broader implications and our federal program is jointly run by DOE and the National Institutes of Health. As the 15-year, $3 billion federal program reached its halfway point this year, the scientific world was stunned on May 9th when one of the country's foremost genetic scientists, Dr. Craig Venter, and the Perkin-Elmer Corporation announced they would form a new venture to, as they put it, "substantially complete the sequencing of the human genome" in 3 years at one-tenth the cost of the federal program.

Just how this should affect the government program is the focus of this hearing today. Press reports and some back and forth between critics and supporters of the federal program have raised as many questions as it has produced answers. For example, are the goals of the initiative realistic or just an optimistic vision? Will this private sector initiative duplicate the federal program and make it redundant or is it another approach that can complement the federal program and make it stronger? Is the pace and the cost of the federal program increased by the bureaucratic nature of any federal program or does the timetable and cost reflect what is necessary to do a thorough job? And will the federal program utilize the latest technology described in the private sector announcement?

Our witnesses today, a cross-section of distinguished scientists from the government and from the private sectors, should be able to supply, I hope, some of the answers to those questions.

One of the witnesses today warns that Congress is the wrong forum in which to debate the relative merits of different scientific approaches to sequencing the human genome. Let me say I couldn't agree more. We're not, as my friend George Brown might say, set up to be a science court.

However, we are given the responsibility of overseeing a federal program that has spent about $1.9 billion to date. The purpose of this hearing is to get the best advice possible on how to—how additional moneys should be spent.

I would also like to take a moment to thank our witnesses for being here today. Some of you traveled long distances at your own expense; others had to rearrange their personal schedules to fit ours, and we certainly appreciate it.

Before I introduce our panel, let me turn to my good friend from Indiana, the distinguished Ranking Minority Member, Mr. Roemer, for his opening remarks.

Mr. ROEMER. I thank our distinguished Chairman and want to applaud him and salute him for this timely hearing on such a complicated, yet fascinating, subject. I would ask unanimous consent that my entire statement be entered into the record, Mr. Chairman.

Chairman CALVERT. Without objection, so ordered.

Mr. ROEMER. And I will just talk for a few seconds and then yield back the balance of my time to this expert panel. Certainly we have heard the mantra in this Congress of faster, cheaper, better. We have heard promises at times from the public sector, and promises at times from the private sector, that appeared too good to be true. Here we have the possibility, a golden possibility, of a private-public partnership that could result in phenomenal return for science and in phenomenal return for the taxpayer. We want to see if these promises, and if this potential, is in fact true and if, in fact, we can do this partnership between the public and private sector that some have talked about. We want to look at the question of privacy and patent issues. We want to look at many other serious questions when it results in cutting the costs as has been talked about in the press by such a significant degree, yet yielding the science that we have been talking about for the last decade. So I'm anxious to hear from our expert witnesses. I'm very, very interested in this topic and we look forward to our expert panel giving us the insight and the advice to fulfill the mantra of faster, cheaper, better, not just with political rhetoric but with real promise for a private sector, public sector partnership. And with that, I yield back the balance of my time.

[The prepared statement of Mr. Roemer follows:]

OPENING STATEMENT
BY
THE HONORABLE

TIM ROEMER

RANKING DEMOCRATIC MEMBER
SUBCOMMITTEE ON ENERGY & ENVIRONMENT

COMMITTEE ON SCIENCE

THE HUMAN GENOME PROJECT: HOW PRIVATE SECTOR DEVELOPMENTS AFFECT THE GOVERNMENT PROGRAM

JUNE 17, 1998
1 PM

I would like to thank the Subcommittee Chairman for his foresight and timely action in calling this hearing. This development is a complicated one, not just in terms of what it will mean for our federal programs, although that is the most prominent question, but in terms of what it will mean for our citizens and our international relationships.

In these times of balanced budgets, tobacco settlements, and huge international projects, the 105th Congress has readily embraced the "faster, better, cheaper" mantra. Often, but not always, for very good reasons. This pattern seems to be holding as we address the decision made by Craig Venter and the Perkin-Elmer Corporation to form a new company that claims it will complete the sequence of the entire genome in 3 years at about 1/10 the cost of the Federal Human Genome Project.

This development has raised the question of whether or not we in Congress should scale back our federal programs based simply on the **promise** of respected and experienced scientists and an equally respected and established private corporation. The purpose of this hearing is to determine if that line of thinking is premature.

At this point, I am more concerned with the inevitable changes that will occur as the mission shifts from public interest to private profit. While I do not discount the sentiment and motive behind the search for this life-saving knowledge, I think that it is only right to address the possible pitfalls of private-sector control of this genetic information. Commercialization can promote the availability of new treatments, but it can also stifle discovery and innovation. Also, issues of privacy have never been fully addressed. The complexity of these issues should not be underestimated and an appropriate balance must be struck.

So I thank you again Mr. Calvert and I welcome our witnesses. I hope that they will be able to shed some light on how the involved parties might form a symbiotic relationship between the Federal Human Gemome Project and the proposed private-sector project, **and** how they plan to ensure that the rights of the American people are not violated or their needs exploited.

Chairman CALVERT. I thank the gentleman.

Our first witness is Dr. Ari Patrinos, Associate Director of Energy Research for the Department of Energy who oversees the human genome project for DOE. Dr. Francis Collins is Director of the National Human Genome Research Institute for the National Institutes of Health; Dr. Craig Venter is President of the Institute for Genomic Research in Rockville, Maryland, and is one of the partners in the private sector initiative announced on May 9th; Dr. David Galas is President and Chief Executive Officer of CHIRO Science R&D Inc. of Washington State. Dr. Galas at one time served as Director for Health and Environmental Research at the Department of Energy; and Dr. Maynard Olson is Professor of Medicine for the Division of Medical Genetics at the University of Washington.

Gentlemen, it's our policy to swear in all witnesses. So I would ask you to rise for me please.

Do you solemnly swear to tell the truth, the whole truth, and nothing but the truth?

Mr. PATRINOS. I do.

Dr. COLLINS. I do.

Mr. VENTER. I do.

Mr. GALAS. I do.

Mr. OLSON. I do.

Chairman CALVERT. You're sworn in. Let the record show that all answered in the affirmative.

You may be seated.

Without objection, the full written testimony for each of you will be included in the record. I would ask that each of you summarize your remarks in approximately 5 minutes so we'll have sufficient time for questions.

Dr. Patrinos, you may begin your opening statement.

TESTIMONY OF ARISTIDES A. PATRINOS, ASSOCIATE DIRECTOR OF ENERGY RESEARCH FOR BIOLOGICAL AND ENVIRONMENTAL RESEARCH, U.S. DEPARTMENT OF ENERGY, WASHINGTON, DC

Mr. PATRINOS. Thank you, Mr. Chairman, Mr. Roemer. I am pleased to testify before the Subcommittee on the future of the human genome project and, specifically, how the new private sector venture, will help shape our program. I'm honored to testify along with such a distinguished set of scientists, the gentlemen to my left. The Department of Energy takes great pride in its pioneering in the human genome project that will essentially revolutionize biology and help usher in a new millennium of wonderful applications in medicine, environmental bioremediation, and sustainable development.

Back in 1986, the Biological and Environmental Research program that I have the privilege of directing presently, while seeking a molecular level understanding of the effects of ionizing radiation on human biology, proposed to sequence the 3 billion base pairs of human DNA and identify the important genes on the 23 pairs of chromosomes.

It was a proposal that at the time was considered with, or at least was met with considerable skepticism and, I might add, some hostility as well. However, the rest is history, as you know, and in 1990, along with our colleagues at the National Institutes of Health, we formally launched the Human Genome Program, along with a common 5-year plan that we updated in 1993 because of faster-than-expected progress. As you mentioned, Dr. Galas, who was my predecessor in this job, was, in fact, in charge of the DOE element of the program at that time. Last month representatives of our two agencies from the NIH and the Department of Energy met with key members of the scientific community to work out the details of the next 5-year plan that we expect to issue in October, officially October of this year, and I expect, we expect that this plan will be coordinated with our international partners such as the Sanger Center in the United Kingdom, as well as with private sector ventures such as initiative that you made reference to, the initiative launched by Dr. Craig Venter of the Institute for Genomic Research and Perkin-Elmer.

At the midpoint of its projected 15-year lifetime, the human genome program is embarking on its high-volume DNA sequencing phase. This has been made possible because of advances in sequencing technologies, because of advances in informatics and also because of enhanced access to cloned resources. The Department of Energy has met this challenge by creating the Joint Genome Institute and merging the resources and capabilities and talents of our three genome centers at our laboratories at Berkeley, Los Alamos, and Livermore. The DOE expects to do its fair share of high-volume DNA sequencing at the sequencing factory that we are establishing at Walnut Creek, California.

From the very beginning the human genome program has focused on developing technologies and resources that would advance the utility and science of the information contained in the human genome and it is in that vein that we welcome the private sector initiatives such as the one announced by Dr. Venter and Perkin-Elmer. That effort is particularly noteworthy because it is our understanding that they will share their data with us promptly, and it also comes at a time when we all collectively recognize that our nation needs enhanced sequencing capacity so that we can all reap the benefits of the human genome project in terms of public health and medicine.

Some of the basic research that the Human Genome Program has nurtured, both at The Institute of Genomic Research and elsewhere, laid the foundation for the sequencing approach that's been proposed by the private sector venture. Such intellectual partnerships between the public and private programs, we believe, will speed the completion of the human genome project goals and significantly enrich the scientific community that's involved in the project. As we speed up the exploitation of the genomic information, however, we should be ever vigilant about the ethical, legal, and social implications that we may have to deal with. During the next few months we will be unveiling the specifics of our new 5-year plan that will definitely incorporate the new private sector venture. The scientific community that is involved in our project is on the cutting edge of technology development and scientific discover, and I have every confidence that many more surprises await us on the road ahead.

I believe that these discoveries will happen at the interfaces between the agencies that are involved in the human genome project such as biology, information science, and engineering, and I think that our program and, from the parochial point of view, our laboratories, the DOE National Laboratories, are ideally suited to contribute to the discoveries for the benefit of our Nation.

This completes my prepared remarks and I'll be ready to answer any questions. Thank you.

[The prepared statement and attachments of Mr. Patrinos follow:]

STATEMENT OF

DR. ARI PATRINOS

ASSOCIATE DIRECTOR

OFFICE OF BIOLOGICAL AND ENVIRONMENTAL RESEARCH

OFFICE OF ENERGY RESEARCH

DEPARTMENT OF ENERGY

BEFORE THE

COMMITTEE ON SCIENCE

SUBCOMMITTEE ON ENERGY AND ENVIRONMENT

UNITED STATES HOUSE OF REPRESENTATIVES

JUNE 17, 1998

Mr. Chairman and Members of the Subcommittee:

I am pleased to testify before the Subcommittee on the future of the Human Genome Project (HGP). The Department of Energy (DOE) takes great pride in its role in this important research endeavor that will revolutionize the field of biology and help usher in a new millennium of wonderful applications in the fields of medicine, environmental remediation, and sustainable development.

The DOE Biological and Environmental Research (BER) program launched a pilot project in 1986 to examine the feasibility of sequencing the three billion pairs of human DNA and to identify all the genes on our twenty-three pairs of chromosomes. One of the initial objectives of the BER project was to seek a molecular-level understanding of the effects of ionizing radiation on human biology, a goal that continues today. The National Institutes of Health (NIH), having started its own program in 1988, joined DOE in the formal launch of the HGP in 1990 and together the two agencies issued a five-year research plan. In 1993, that plan was updated two years ahead of schedule, due to faster than expected progress; most notably, rapid progress came from advances in physical mapping and in technology, and simultaneously from the unexpected pace of disease gene discovery that dramatically demonstrated the value of genome-scale research. Last month, representatives from the two agencies met with key members of the scientific community to agree on the details of the next five-year plan that will be released in October 1998. The plan will be coordinated with those of our international partners (e.g., with the United Kingdom's Sanger Center) as well as with parallel private sector initiatives such as the recently announced venture by Perkin-Elmer and Dr. Craig Venter of The Institute for Genomic Research (PE-TIGR).

At the midpoint of its projected 15-year lifetime, following achievement of every milestone of the 1993 plan on or ahead of schedule, the HGP is embarking on the task of high volume human DNA sequencing in order to deliver the highly accurate sequence of an entire generic human genome by 2005; the task has been made possible by advances in sequencing and information technologies and in enhanced access to clone resources. The DOE has responded to the new challenges of this phase of the HGP by creating the DOE Joint Genome Institute (JGI), the combination of the DOE genome research centers at Los Alamos, Lawrence Berkeley, and Lawrence Livermore National Laboratories. The Institute will undertake the DOE's share of high volume sequencing at its new production sequencing facility in Walnut Creek, California.

The new five-year plan will describe the details of the public sector sequencing strategy as well as the other elements of the HGP. In addition to the pursuit of a complete map of the human genome, these elements include: the further development of sequencing technologies that will be needed to use information being generated in the HGP long after the first human sequence is completed in 2005; the creation of the data bases that will accept and process the large amounts of data generated by sequencing; the sequencing of genomes of model organisms to help us understand, most efficiently and cost effectively, the human genome; the ethical, legal, and social implications (ELSI) of the HGP; and the pursuit of some of the biological applications that will be enabled by the completion of the first reference or generic genome sequence, a sequence comprised of DNA from ten women and ten men who will be rigorously anonymous and whose informed consent will have been fully assured.

Progress in the HGP itself, together with scientific contributions from the many HGP spinoffs in both the public and private sector, will enable us to include new program goals that could not have been anticipated only a few years ago. These unexpected new goals are consistent with the history of the HGP making bigger payoffs and providing even greater value than anticipated, both scientific and economic. Advances in technology will enable the efficient characterization of the biological functional units in every cell, the gene transcripts and their protein products. Moreover, rapid progress in determining the genomic sequences of model organisms such as yeast (the first yeast genome was completed in 1996), the worm, C. Elegans, (scheduled for completion in 1998), and a rapidly increasing number of microbes is enabling more rapid characterization and discovery of human genes than previously expected. Progress in meeting the sequencing and biological goals of the HGP will also challenge the ELSI component of the HGP to address, more quickly, the critical issues arising from the unexpectedly rapid availability and use of human genome information.

From the beginning, the HGP has been focused on developing technologies and resources that would advance the science and utility of the information contained in the human genome. Thus, DOE welcomes private sector initiatives such as the PE-TIGR venture that will add value to the public sector effort. This private sector effort is particularly noteworthy since it is our understanding that PE-TIGR intends to share its data promptly with the HGP, and since it comes at a time when there is an increased need for sequencing capacity if the Nation is to realize fully the public health and medical benefits of the genome project as quickly as possible.

It is notable that NIH- and DOE-funded basic research (at TIGR and elsewhere) laid the foundation for the sequencing approach being proposed by PE-TIGR. We do believe that such emerging public-private intellectual partnerships will speed completion of some HGP goals and enrich the scientific community involved in the HGP. However, at the same time, it is important that we work to guarantee that HGP data acquired with public funds continue to be made available to the scientific community at large and that the data is of a quality that provides the greatest scientific information and utility. The product of the PE-TIGR venture will contain many gaps, whereas the HGP has always been committed to a contiguous, high quality, highly accurate, complete sequence. Moreover, there is a critical need for increased sequencing capacity within our academic and national laboratories to meet the many public sector sequencing demands that will follow the HGP. This information will be revealed by sequencing the genomes of model organisms, such as mice, rats, and primates for which we have a rapidly growing wealth of biological information that provides insight into how human genes function. In addition, sequence information from portions of the genomes of hundreds of individuals will be needed to understand human genetic variation and will serve as the basis for developing individual-specific diagnosis and therapy, a potential focus of 21st Century medicine.

The scientific community involved in the HGP is truly on the cutting edge of technology development and scientific discovery; and as a result, surprising new discoveries and advances can be expected over the next few years. Many of these discoveries will occur at the interfaces of the sciences that are involved in the HGP such as biology, information science, and engineering. The multidisciplinary capabilities of our national laboratories are ideally suited to contribute to these discoveries. Together with our NIH partners we strive to facilitate these discoveries and advances for the benefit of the Nation.

This completes my prepared testimony. I would be happy to answer your questions.

Ari Patrinos

Dr. Patrinos received a diploma in mechanical and electrical engineering from the National Technical University of Athens and a PhD in mechanical engineering and astronautical sciences from Northwestern University. His research included atmospheric turbulence, computational fluid dynamics, and hydrodynamic stability. After a year on the faculty of the University of Rochester he joined Oak Ridge National Laboratory in 1976 to conduct research on energy-related weather and climate modification and to develop numerical codes for loss-of-coolant (LOC) nuclear accident simulations as well as for river flows and lake circulations.

In 1980, he joined Brookhaven National Laboratory to develop atmospheric chemistry models and to lead field programs on wetfall chemistry. In 1984, he was detailed to EPA and to the National Acid Deposition Assessment Program (NAPAP) staff in Washington, DC. He joined DOE in 1986, restructuring the Department's atmospheric sciences program, and in 1988 led the expansion of DOE's research effort in global environmental change. He was the director of the Atmospheric and Climate Research Division (ACRD) of DOE's Office of Biological and Environmental Research (OBER) until 1990. When ACRD was merged with OBER's Ecological Research Division, he became director of the combined Environmental Sciences Division.

From August 1993 until March 1995, Dr. Patrinos was acting as the Associate Director for Biological and Environmental Research in the Office of Energy Research; since March 1995 he has been the Associate Director, who oversees the research activities including the DOE human and microbial genome programs, structural biology, nuclear medicine and health effects, global environmental change, and basic research underpinning DOE's environmental restoration effort. Dr. Patrinos represents DOE on several subcommittees of the Committee on Environment and Natural Resources of the National Science and Technology Council. He is a member of the American Society of Mechanical Engineers, the American Geophysical Union, the American Meteorological Society, and the Greek Technical Society.

Chairman CALVERT. Dr. Collins.

TESTIMONY OF FRANCIS S. COLLINS, M.D., DIRECTOR, NATIONAL HUMAN GENOME RESEARCH INSTITUTE, NATIONAL INSTITUTES OF HEALTH, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES, BETHESDA, MD

Dr. COLLINS. Thank you very much, Mr. Chairman. I am honored to appear before this Committee, especially with the distinguished folks sitting at the table with me. I am Director of the National Human Genome Research Institute which is the part of the National Institutes of Health which is devoted to the human genome project, one of 22 such institutes and centers of the NIH.

In case you are not familiar with the NIH's means of funding science, let me just quickly point out that the funding that we give to the Human Genome Project is derived from grant applications which we get from investigators at universities, institutes and some companies around the country. They send in their grant proposals to us. Those are peer reviewed and then we select the ones that we think are the most meritorious for funding. Regrettably at the present time, only about one in four approved applications is funded but that is where the work of the NIH component of the genome project is done, out there in academia, in small companies, and in institutes.

I wanted to make four points in my brief opening statement which are taken from the written remarks which are more extensive. First of all, Mr. Chairman, you pointed out that there have been bold words spoken about the genome project. Let me speak a couple of them myself. As a physician and a scientist, I do believe that genetics has become the core science of medicine. Whatever disease you're interested in understanding, genetics is now the most powerful tool you have to get at the mysteries that still remain unlocked. I also believe that the genome project has become the center of genetics, this effort to map and sequence all the DNA of the human and other model organisms is very much the focal point of the modern revolution. So what we are talking about today is the core of the core. Its importance can hardly be overstated. I do believe historians will look at this as the most ambitious and important organized scientific effort that humankind has mounted, including splitting the atom or going to the moon, because this is an investigation into ourselves.

Second point: The genome project has been characterized by a complex, but carefully planned, agenda since the outset. There has been some misunderstanding I believe, and perhaps recently especially in the press, about what the genome project aims to do. This is not just a project to sequence human DNA. In its first several years, many of the goals of the project related to developing maps, genetic maps and physical maps, as well as improving the technologies in order to be able to afford to do the human sequencing at the pace that was needed to complete the job at the cost that was estimated to be available. So up until now, in fact, only a minor fraction of the budget of the human genome project has been devoted to the actual human sequencing, the part that is now ramping up in a major way with 10 percent of that now available in public database in assembled or partially assembled form.

There is also an emphasis on model organisms which has taught us much about how genetics predicts a particular kind of phenotype and which will serve us well in trying to understand what the human DNA sequence means. And there is our ELSI program which Dr. Patrinos has already mentioned, looking at the ethical, legal, and social implications of this research. So the genome project is much broader than just the human sequence. When we look at cost comparisons, for instance, of this approach versus that approach, it would be important to be sure we are talking about the same activities.

Third point: The genome project up until now is arguably one of the more impressive success stories of the federal investment in science of all time. Every milestone that has been put forward by carefully chosen advisers outside the government have been achieved or exceeded. The cost that has gone into this project is roughly 25 percent less in its first half than was expected by the original planners, so it is fair to say the project has been faster, better, and cheaper up until now and we aim to maintain that record.

As a physician I can tell you the consequences of this project are all around us. Back in the 1980's, when I was on the faculty at the University of Michigan, I spent almost 10 years finally identifying the cystic fibrosis gene and another roughly 10 years participating in a group that found the Huntington's disease gene. That was the best you could do in the 1980's. Nowadays, it's a matter of months. Just a few months ago, a gene for Parkinson's disease was found, using the tools of the genome project, in 9 months, and breaking open research in that field which has really been frustrating for 30 years. So this is a success already. You don't have to wait until the sequence is in hand to see it happen.

Fourth point: Partnership with the private sector is both necessary and desirable and we welcome this new initiative which is being discussed today by Dr. Venter. In fact, such public/private partnerships have characterized the genome project from the outset. There are many other examples of that sort, though perhaps none as bold as this one. Again, we need to look carefully at the ways in which this private initiative and the publicly-funded effort can be complementary and we also need to consider scientifically the ways that the strategy is different, which actually adds to the complementarily. And I know Dr. Olson will particularly comment upon that in his remarks.

Let me assure you, we will work together. If you doubt that, notice that Dr. Venter and I seem to have worn the same clothes today without intending to. We are intending to be partners in this in every possible way, so let this be a symbol thereof.

This is not a race. We will work together, we believe in the value of that, we believe we have complimentary strategies. The federal effort is fully prepared to adjust their strategy. As we move forward we have a vigorous advisory process to do that, constituted by some of the world's best scientists. We have adjusted our strategy on a regular basis, based on technological developments, but I would argue that it's a little soon to know exactly what that adjustment should be. As Dr. Venter will tell you, the proposal which has been put forward is bold, but is yet untried, and the quality of the product, a very serious question because we do believe we want the whole genome sequence with as few gaps as possible, as few mistakes as possible, the quality is so important that one must not, I think, deviate from that goal or from the strategy to get there until we have the data in front of us to see how this new approach will work.

In that regard, we welcome a proposal by Dr. Venter to try out, as a pilot effort, the genome sequence of the fruitfly Drosophila. This effort, which will get under way in about 6 months, focuses on an organism whose genome is 30 times smaller, and much more tractable and I believe we will learn a lot from that pilot effort about the ways in which this strategy can be applied to the human. At that point it will be easier, perhaps, for the federal effort to make some predictions about ways that we might adjust our strategy.

But to summarize, we welcome this development, we believe that we have a good track record of working together with the private sector, and I look forward to seeing these two complimentary efforts get us there soon, which is my goal, and should be yours.

[The prepared statement and attachments of Dr. Collins follow:]

National Institutes of Health

Statement of

Francis S. Collins, M.D., Ph.D.

Director, National Human Genome Research Institute

on

The Human Genome Project:

How Private Sector Developments Affect the Government Program

before the

Subcommittee on Energy and the Environment

Committee on Science

Unites States House of Representatives

June 17, 1998

I am Dr. Francis Collins, Director of the National Human Genome Research Institute (NHGRI) of the National Institutes of Health. I appreciate the opportunity to appear before the Subcommittee today to discuss the Human Genome Project and the implications of the recent announcement by a private company of their intentions to carry out large-scale sequencing of the human genome.

The NHGRI is one of the 22 Institutes and Centers that comprise the federation of federal research entities known as the National Institutes of Health (NIH). The vast majority of research dollars appropriated to the NIH flow out to the scientific community across the Nation, primarily in the form of peer-reviewed research grants. Today, that community numbers more than 50,000 investigators affiliated with nearly 2,000 universities, hospitals, and other research facilities located in all 50 states, the District of Columbia, Puerto Rico, Guam, the Virgin Islands, and certain points abroad.

The NHGRI is the lead Institute at the NIH with responsibility for The Human Genome Project (HGP). The HGP officially began in October of 1990 as a 15-year program to characterize in detail the complete set of human genetic instructions (the "genome"). The central aim of the project, which the federal government funds through programs at the NIH's National Human Genome Research Institute and the Department of Energy, is to arm health researchers with powerful gene-finding and DNA analysis tools to unravel and understand the myriad human diseases that have their roots in DNA. Now at its half-way mark, genome project tools have underpinned virtually all gene discoveries of this decade.

The Human Genome Project's success stems largely from a unique and rigorous planning process that sets ambitious research goals, time lines and budgets. The first joint NIH/DOE plan, which covered years 1991-1995, included goals for:

- physical and genetic maps;
- experimental DNA sequencing of the fruit fly, a round worm, yeast, and the bacterium *E.coli;*
- computer management of research data; and
- studies of the ethical, legal, and social implications (ELSI) of these new abilities to read genetic information

Because of the rapid pace of genome research and technology development, scientists met many of those initial goals ahead of schedule and under budget. So the research plan was updated again in 1993 to establish new NIH-DOE goals through 1998. All of these goals have now been met or exceeded. Original expectations were that the NIH cost of these activities from FY'91-97 would exceed $1 billion in 1991 dollars. I am pleased to report that the cost has been about 25 percent less than that projection.

Gene Discovery

Today, with Human Genome Project tools, it is possible to track down a disease-related gene even when nothing is known about the biochemical problems of the disease or how the gene works. This technique, based on identifying the position of a gene in the chromosome and then isolating it, is commonly referred to as positional cloning and was successfully used for the first time in 1986. Now, the increasing detail and quality of genome maps have reduced the time it takes to find a disease gene from years, to months, to weeks, to sometimes just days, and scientists are using the tools to discover dozens of disease genes each year.

An Example - Parkinson's Disease

The isolation of a gene for Parkinson's disease (PD) last year demonstrated the power of this new discovery method and showed conclusively that changes in DNA can cause PD in some families. Only two years ago, the National Institute of Neurological Disorders and Stroke held a workshop to explore using genetic approaches to understand PD. A team led by scientists in NHGRI's Division of Intramural Research (DIR) began large-scale genetic analysis of DNA from members of a large Italian family containing almost 600 people, more than 60 of whom have been diagnosed with Parkinson's. In nine days, NHGRI gene hunters mapped the gene to a region of chromosome 4, which contained approximately 100 genes. One of the several genes in that interval had already been identified on the gene map and was known to encode a protein called alpha-synuclein.

In just a few months, the researchers showed conclusively that an altered alpha-synuclein gene caused Parkinson's disease in the study families. Many have hailed this as the most significant advance in Parkinson's disease research in 30 years. Just last month, a Japanese research team used genome mapping tools to isolate another gene, this time on chromosome 6, that also appears to contain a gene that, when altered, predisposes the individual to a rare juvenile form of Parkinson's disease.

Ethical, Legal, and Social Implications

NHGRI has established productive partnerships among consumers, scientists, and policy makers to help reduce the possibility that genetic information will be used to harm an individual or family members and ensure that it will be of benefit to both patients and providers. As an integral part of the Human Genome Project, the NHGRI and the DOE have each set aside a portion of their funding to anticipate, analyze, and address the ethical, legal, and social implications (ELSI) of the Project's new advances in human genetics. The current goals of the ELSI program are to improve the understanding of these issues through research and education, to stimulate informed public discussion, and to develop policy options intended to ensure that genetic information is used for the benefit of individuals and society. Because genetic information is personal, powerful, and potentially predictive, it can be used to stigmatize and discriminate against people. Genetic information must be private.

DNA Sequencing

If the letters representing the 3 billion bases in the human genome were printed out in books, and the books were stacked one on top of the other, they would reach as high as the Washington Monument. The current major goal of the Human Genome Project is to read the order, letter by letter, of those 3 billion bases.

Sequencing was once done by hand as a series of chemical reactions—a slow and costly method. In 1990, when the HGP began, the sequencing cost was $10/base. Now, because of public investment and collaboration with the private sector, machines read the sequence fragments quickly and efficiently. As a result, the sequencing cost has been dramatically reduced to roughly $.50/base for high-quality "finished" sequence.

Using a strategy referred to as "shotgun" sequencing, an investigator takes each page of those books stacked as tall as the Washington Monument, and randomly cuts the text into small fragments. These fragments are small enough for sequencing machines to read. To get long stretches of contiguous DNA, investigators must then reassemble these sequenced fragments back into sentences, paragraphs, chapters, and books. The reassembly of this puzzle is carried out largely by sophisticated computer programs.

The sequencing strategy the public genome project uses employs shotgun sequencing of DNA fragments that already have been carefully mapped and catalogued. This process makes reassembling the sequenced fragments into contiguous sequence easier because you know where the fragment came from. In addition, scientists periodically encounter DNA fragments that are particularly difficult to sequence. To return to the analogy, it is much easier, takes less time, and is less costly to assemble the text in "finished" form if all the fragments are known to have come from the same chapter.

In 1996, NHGRI began pilot projects to test strategies and technologies for full-scale sequencing of the human genome. We now have undertaken human sequencing in earnest. As a result, investigators have deposited almost 150 million bases of "finished" high-quality human DNA sequence in GenBank, the publicly funded database supported by the National Library of Medicine. In accordance with the agreed-upon standards of the international genomic community, all NIH-DOE funded sequencers have agreed to a rapid data release policy, such that, new sequence data is submitted to publicly accessible data banks within 24 hours. If one includes "finished" and "close-to-finished" sequence, over 300 million bases, or 10 percent, of the human DNA sequence has been deposited in GenBank.

In order to meet the standards adopted by the international genomic community, the sequence produced must have four characteristics --the "4 A's" of the Human Genome Project --

1) the sequence must be **accurate**, that is, the DNA spellings must be correct. The publicly funded genome effort will ensure accuracy of 99.99 percent or better.

2) the sequence must be **assembled**. Large-scale sequencing relies on the accurate assembly of smaller lengths of sequenced DNA into longer, genomic-scale pieces, so DNA will be assembled into long pieces that reflect the original genomic DNA.

3) Because human DNA sequence must also be **affordable**, a portion of our research funds focuses on technology development to reduce the cost as much as possible.

4) Finally, high-quality, finished human DNA sequence must be **accessible.** In order to be useful, sequence data needs to be rapidly available to the entire research community.

Research Planning

Informed by a series of workshops over the past year that reviewed research progress and identified genome research opportunities, Human Genome Project leaders recently met with more than 100 representatives from a range of scientific disciplines to develop the next 5-year plan, scheduled to begin in the fall of 1998. With both the physical and genetic maps complete, and human DNA sequencing pilot projects underway, goals of the 1998-2003 draft plan considered at that meeting focused on:

- completing a full, highly accurate and contiguous human genome DNA sequence;
- further development of technologies for steadily increasing sequencing capacity and reducing costs;
- studies of variations in human DNA;
- studies of how large sets of genes function;
- studies of the similarities and differences between the human genome and those of important laboratory animals;
- improved computer methods for data management; and
- studies regarding the ethical, legal and social implications of the HGP.

Private Sector Developments

Just prior to the HGP planning meeting, industry researchers from The Institute for Genomic Research (TIGR) and Perkin Elmer, Inc. announced a plan to apply a DNA sequencing strategy they had used on micro-organisms to produce a "rough draft" of the human genome sequence. The sequencing strategy recently proposed by Perkin-Elmer, Inc. and TIGR differs from the public effort in two significant ways: quality and access.

First, that strategy, called "whole-genome shotgun sequencing", employs fragments that have not been previously mapped or catalogued prior to sequencing. Because scientists will not know where in the long chain of 3 billion base pairs the fragment might belong, the task of reassembling the fragments becomes far more difficult. This difficulty in reassembly inevitably will lead to gaps and misassemblies in the sequence. Some of these may occur in DNA regions with great biological significance. The private sector approach does not propose to fill in all the gaps left by these unsequenced fragments, thereby creating a product that will be incomplete for many research uses.

Secondly, release of sequence data from the Perkin-Elmer-TIGR effort will occur quarterly, rather than daily. The policy of daily release of DNA sequence data by publicly-funded efforts was arrived at because of the great interest in the scientific community in gaining access to this highly valuable information. Any delay can result in wasted effort in research.

Deliberations on Five-Year Research Plan

Because the industry plan seemed to parallel some aspects of the federal Human Genome Project, planners and advisors to the NIH-DOE program have been debating extensively how the two proposals could be matched up. The scientists, at the recent planning meeting on the draft HGP 5-Year Plan, concluded that while the two projects should complement one another, the federal project should continue its plans to provide high-quality human DNA sequence as soon as possible and that all data should be freely accessible.

Those conclusions rested on a few key factors:

- The industry effort may not deliver the product in the time and manner proposed. The industry approach to sequencing has not been tried on large and complex genomes, such as the human, and depends on newly developed and unproven machines. Data to evaluate the "whole genome" shotgun approach will initially come from a trial project on the fruitfly, Drosophila, but is not expected on the human for at least 12 to 18 months;
- The industry plan will produce a large amount of highly useful sequence data, but this plan will yield a qualitatively different product that will likely contain tens of thousands of gaps;
- The industry plan calls for release of sequence data on a quarterly basis, and patenting of 100-300 "gene systems." While quarterly data release is commendable, the plan is not as strong as the standards established by the international sequencing community which require release of data within 24 hours and discourage patenting. Further, some concerns were expressed that the private effort's commitment to data release might diminish over time, if business pressures came to the forefront.

In view of those concerns, advisors at the planning meeting enthusiastically made several unanimous recommendations:

- The publicly funded genome project should continue with plans to provide a complete, high-quality human DNA sequence by the year 2005, and sooner if at all possible;
- All possible steps must be taken to ensure that all sequence data remain in the public domain;

- The publicly funded effort should take advantage of technology advances to increase sequencing capacity as much as possible as soon as possible to meet research needs, both for sequencing of the human and model organisms; and

- The sequencing of DNA regions of high utility and research interest should be emphasized.

Now, Human Genome Project leaders at the NIH and DOE are considering that advice as they put the final touches on the new research plan, which will be published in the fall of 1998. The complete plan will contain details for all of the Human Genome Project's goals, including sequencing, gene function, human variation, technology development, and Ethical Legal and Social Implications.

The private and public genome sequencing efforts should not be seen as engaged in a race. In fact, scientists at TIGR and Perkin-Elmer have expressed their enthusiasm for a continued vigorous public effort on the HGP, and have conveyed their willingness to collaborate with NIH and DOE on the production of the complete human sequence. The NIH and DOE welcome this collaborative approach, as the whole should be greater than the sum of the parts.

Conclusion

Mr. Chairman, I commend you, and the Members of this Subcommittee, for convening this hearing today. The impact on the future of biology of knowing the order of all 3 billion human DNA bases has been compared to Mendeleev's establishment of the Periodic Table of the Elements in the 19th century and the advances in chemistry that followed. The complete set of human genes--the biologic periodic table--will make it possible to begin to understand how they function and interact. Rapidly evolving technologies, comparable to those used in the semi-conductor industry, will allow scientists to build detectors that analyze tens of thousands of genes in a single experiment. Scientists will use the powerful new tools to reveal the secrets of disease susceptibility. This knowledge will in turn allow researchers to create broad new opportunities for preventive medicine, lay the foundation needed to develop and better target effective therapeutics, and provide unprecedented information about the origin and migration of human populations.

The investment of substantial funds by the private sector in human sequencing reaffirms the enormous value of Human Genome Project products and is a testament to the success and value of the tools already developed by the publicly supported project. For the reasons outlined above, it is not yet known what role this new endeavor will play over the long term in providing the publicly available, detailed "A-to-Z" instruction book ultimately promised by the Human Genome Project. Project leaders at the National Institutes of Health and the Department of Energy look forward to close cooperation with Perkin-Elmer and TIGR as the new initiative unfolds over the next few years.

This concludes my remarks. I would be pleased to answer any questions.

Francis S. Collins, M.D., Ph.D., Dr. Francis Collins was appointed Director of the National Human Genome Research Institute in April 1993. NHGRI oversees the role of the National Institutes of Health in the U.S. Human Genome Project.

Dr. Collins pioneered the development of a powerful gene-finding method known as "positional cloning," which utilizes the inheritance pattern of a disease within families to pinpoint the location of the gene associated with the disease. Positional cloning is now commonly used to isolate genes even when no information about the gene's function or biochemistry is known. Dr. Collins is perhaps best known for using positional cloning techniques to isolate the genes for cystic fibrosis, neurofibromatosis type 1, Huntington's disease, and ataxia telangiectasia.

He was formerly a Howard Hughes Medical Institute investigator and professor in the Departments of Internal Medicine and Human Genetics at the University of Michigan School of Medicine in Ann Arbor. He was also director of the NCHGR-supported human genome center at Michigan.

Current active research projects in the Collins laboratory include the develop of better methods for analyzing mutations in disease genes, especially for the BRCA1 gene on chromosome 17. The laboratory is also involved in an ambitious effort to map the major genes contributing to adult-onset diabetes, by carrying out extensive linkage analysis on affected siblings, largely collected in Finland. Positional cloning of the genes for familial mediterranean fever and multiple endocrine neoplasia are also underway, in collaboration with other investigators.

Born in Staunton, Virginia, in 1950, Dr. Collins received his bachelor of science degree with highest honors from the University of Virginia. He received both his M.S. and Ph.D. degrees in physical chemistry from Yale University and an M.D. degree from the University of North Carolina School of Medicine. He completed his internship and residency in internal medicine at the North Carolina Memorial Hospital. From 1981 to 1984, he was a fellow in human genetics and pediatrics at Yale. He joined the Departments of Internal Medicine and Human Genetics at Michigan in 1984, becoming professor in 1991. He became a Howard Hughes Medical Institute assistant investigator in 1987 and full investigator in 1991. Collins is a diplomate of the American Board of Internal Medicine, the American Board of Medical Genetics, and the American College of Medical Genetics.

Dr. Collins was elected to the Institute of Medicine in 1991 and the National Academy of Sciences in 1993. He is also a member of the American Federation for Medical Research, the American Society for Clinical Investigation, the Association of American Physicians, and the international Human Genome Organization. He serves as an associate editor for several publications, including *Genomics*; *Genes, Chromosomes and Cancer*; *Human Molecular Genetics*; *Somatic Cell and Molecular Genetics*; and *Human Mutation*.

Among his most recent awards and honors, Dr. Collins has received the Gairdner Foundation International Award, the Young Investigator Award of the American Federation for Clinical Research, the Doris Tulcin Award for Cystic Fibrosis Research, University of Michigan's Distinguished Faculty Achievement Award, the National Medical Research Award, and the University of Pittsburgh Dickson Prize. He holds honorary degrees from several academic institutions.

Chairman Calvert. Thank you, Doctor.
Dr. Venter.

TESTIMONY OF J. CRAIG VENTER, PRESIDENT AND DIRECTOR, THE INSTITUTE FOR GENOMIC RESEARCH, ROCKVILLE, MD

Mr. Venter. Thank you very much, Mr. Chairman. I appreciate the opportunity to testify before your Subcommittee about the impact our new developments on the federally-funded human genome effort. I also appreciate the comments of Dr. Patrinos and Dr. Collins.

I'm the founder and President of The Institute for Genomic Research, often known as TIGR, in Rockville, Maryland, and I'm the to-be President of the new company we're forming, I'm a co-founder of that company along with Tony White and Mike Hunkapiller of the Perkin-Elmer Corporation. Recent publicity about our new venture to sequence the human genome in 3 years has lead to speculation that funding for the human genome effort should be reduced or eliminated. Nothing could be further from the truth. Upon completion of today's hearing, I hope it's clear that this new private venture, and the federally-funded project are, in fact, complimentary efforts that can work together to make unprecedented impact on improving research on human health.

One goal of our new to-be-named company is to sequence the human genome over 3 years, using dramatic new technology developed by Mike Hunkapiller's team at the Perkin-Elmer Corporation in strategies that have been developed by myself and my colleagues at The Institute for Genomic Research for sequencing whole genomes. I agree with the comments of Dr. Collins that the focus has been lost in the purpose of obtaining the human genome sequence. And it was concentrating on what was perceived to be an absolutely monumental task of obtaining that sequence, due to the limits and technologies and procedures that we've had in the past. Analogies to the Manhattan Project and Apollo Project are often used. Billions of dollars from the U.S. Government and Europe and Japan, decades of work from thousands of scientists around the world, were thought to be required to obtain that sequence. New technologies and strategies now change and replace some of these assumptions. The human genome will be accurately and completely covered in one facility by a new company in Rockville, Maryland, with a few hundred workers using new technology.

Our effort has been described by some as a rough draft or worse of the human genome but I've heard these comments before in 1994 when Nobel Laureate Ham Smith and I proposed the new strategy for sequencing genomes. In fact the first genome in history that we published in *Science* in 1995 was done with this approach. The genome review panel involving NIH funding rejected our grant as being impossible and that we'd have a large number of noncloseable gaps and misassembled pieces of the genome and at the best the sequence would be an incomplete and full of holes. They were clearly wrong.

TIGR is the only organization in the world to have completely sequenced more than one genome. In fact, we've completed seven, including the first three and those seven represent half of the entire world's complement of completed genomes. All seven, plus five more to be finished this year by us, were done by the whole genome shotgun approach. Our sequences are some of the highest-quality sequences ever completed and published. More than a dozen pathogen genome projects are now under way at TIGR, including the malaria genome with funding by the National Institutes of Health. I should point out that the Department of Energy using slightly different review processes funded TIGR to sequence two out of three of the first genomes completed in history and that funding was obtained prior to the completion of the Hemophilus influenza sequence in 1995.

The DOE has also funded TIGR to sequence more than a dozen key environmental genomes, using the whole genome shotgun method, and the Department of Energy has also funded the bacterial artificial chromosome in sequencing strategy that is providing the scaffolding for assembling the entire human genome sequence. I'm here to urge you not only to not cut the DOE or other genome budgets because of our announcement and effort, but to actually consider increasing it.

Having the complete genome moves forward all the issues associated with genomics. The sequence is the beginning of the genome project. It is absolutely not the end of anything, except, perhaps, the end of ignorance. A private/public partnership will not only ensure completion of the genome sequence sooner, it will provide the basis for beginning the key aspects of the genome project, for example, understanding what the sequence means.

Because our effort is moving forward substantially the timetable for completing the genome sequence, the resources for understanding the genomic code become even more important. With comparative genomes, we've learned this in microbial genome sequences, having one genome was fantastic, having two or three was phenomenal and aided our understanding. That's the situation with human and that's part of the existing plan to do the mouse and other genomes. We need those genomes to understand and interpret the human genome. By working together, DOE, NIH, and other public and private institutions can help meet the goal of having a complete map and sequence of the human genome within three years. I see that as an announcement that everybody can be proud of.

I hope that after this hearing you will view our announcement in the federal program, for which you are responsible, not as an either/or proposition, but instead will focus on how these two activities, working in tandem, can ultimately improve our lives and those of generations to come.

This concludes my remarks and I'm pleased to answer any questions you may have.

[The prepared statement and attachments of Mr. Venter follow:]

9712 Medical Center Drive, Rockville, Maryland 20850
(301) 838-0200
(301) 838-0208 Fax

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

PREPARED STATEMENT OF

J. CRAIG VENTER, Ph.D.

PRESIDENT AND DIRECTOR

THE INSTITUTE FOR GENOMIC RESEARCH

BEFORE THE

SUBCOMMITTEE ON ENERGY AND ENVIRONMENT

U.S. HOUSE OF REPRESENTATIVES COMMITTEE ON SCIENCE

June 17, 1998

Mr. Chairman, I appreciate the opportunity to testify today before your subcommittee about the impact of private sector developments on the federally-funded Human Genome Project. Recent publicity surrounding the intent announced by Perkin-Elmer and me to sequence the human genome has led some to speculate that federal funding for the human genome is no longer needed. Nothing could be further from the truth. The Human Genome Project is truly a success that both the scientific community and the federal government can look upon with pride which will continue to generate important information. I am pleased to be here to put in context the role that I have played up until now, and the role the I hope to play in the future. I hope after today you will recognize the success of the program that you have funded, and also recognize the vast potential to improve human health that lies just around the corner by linking both the federally-funded initiative and our new private sector venture.

I am J. Craig Venter, President and Director of The Institute for Genomic Research (TIGR), an independent, not-for-profit research institute in Rockville, MD that I founded in 1992 after leaving the National Institutes of Health (NIH). On May 11, The Perkin-Elmer Corporation, the largest producer of DNA sequencing technologies in the U.S., and I announced a new venture to create a company that will sequence, as part of its initial projects, the *Drosophila* (fruit fly) genome and the human genome within the next three years. These two sequencing projects will be undertaken using breakthrough DNA sequencing technology developed by Perkin-Elmer, and a DNA sequencing strategy that was pioneered by my colleagues and me at TIGR, known as the whole-genome shotgun sequencing method.

This announcement is very exciting for both the public and private scientific communities throughout the world, but it is of particular significance to the United States because it is the validation of the scientific claims of the Human Genome Project, that was first discussed over 14 years ago and funded for the last ten years by U.S. taxpayers. However, I believe that in order for me to explain this comment and adequately answer the question that is the reason for today's hearing, it is important to discuss the events that made our announcement possible.

NIH, ESTs, AND TIGR

When I was at NIH, I was a Section Chief at the National Institute for Neurological Disease and Stroke (NINDS). My lab was involved in a large scale chromosome sequencing effort to discover genes associated with neurological functioning and disease. During this research, my colleagues and I developed a new strategy for identifying genes more rapidly and at much less expense than previously had been possible. Prior to the development of this new strategy we had labored for many years using "traditional" sequencing methods to identify a few genes. In my own case , I spent ten years on the gene for the adrenalin receptor. With the new strategy we greatly exceeded the work of many previous years of effort in just a few months. This new strategy known as Expressed Sequence Tags (ESTs) was published in the journal *Science* in June 1991

(Complementary DNA Sequencing: "Expressed Sequence Tags" and the Human Genome Project. *Science* 252, 1651-1656 (1991)). At the time of this publication, fewer than 2,000 of the 60,000 to 80,000 human genes were known.

It is important to note that this new strategy was more than just creative thinking on the part of the federally-funded scientists in my lab. It also included a significant role played by a new technology company with which we had begun to collaborate. In the late 1980's, Applied Biosystems manufactured a new DNA sequencing technology that greatly improved the speed with which a DNA sequence could be obtained. My NIH lab entered into a CRADA with this firm and worked with them to improve their technology. In fact, this was the first CRADA entered into by NIH with a commercial organization.

By linking my lab's new EST strategy with Applied Biosystem's sequencing technology it became possible to greatly improve the speed with which new genes and DNA sequences in general could be identified. While our new strategy was not yet widely accepted, I learned that orders for the Applied Biosystems DNA sequencers that we used in our experiments had skyrocketed. So there was clearly significant movement on the part of both academic and commercial institutions to adopt this new technique detailed in the *Science* publication.

About a year earlier, Congress had provided the initial funding to the Department of Energy (DOE) and NIH for the Human Genome Project (HGP). From its inception, major technical innovations were considered essential to the success of the project and our new strategy was a significant step forward. In fact, the gene discovery phase of the project could be shortened to almost one-tenth of the originally anticipated timeframe. However, there were many other hurdles to clear.

Obviously with this exciting new strategy I was eager to scale up our research program at NIH in order to implement a successful, large-scale genome sequencing and gene discovery program. However, the extramural genome community did not want genome funding being used on intramural programs. In addition, there was growing controversy surrounding the issue of the U.S. government patenting ESTs that I discovered. I was frustrated that I would be unable to participate in the revolution in biology that we had helped start. I did not want to leave NIH, but after much soul-searching I felt it was the most appropriate option.

In 1992, with funding from the venture capital community, I formed TIGR as an independent, not-for-profit research institute to implement the programs that I had envisioned for my lab at NIH. In short order, we utilized the EST strategy to identify more than half of the genes in the human genome and published this information in the Human Genome Directory in the journal *Nature* in 1995 (Initial Assessment of Human Gene Diversity and Expression Patterns Based Upon 52 Million Basepairs of cDNA Sequence. *Nature* 377 suppl., 3-174 (1995)). Also in 1995, using a new strategy for DNA sequencing that we pioneered, known as the whole-genome shotgun approach, TIGR published the first complete sequence of a self-replicating, living organism, *Haemophilus* *influenzae*, a bacteria that causes ear infections in children (Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science* 269, 496-512 (1995).

In the time since then, TIGR has become one of the leading genomics institutions in the world by determining the complete DNA sequence for six other organisms. Most recently, we published the sequences for the pathogen that causes Lyme disease, *Borrelia burgdorferi*, and the bacteria that causes stomach ulcers, *Helicobacter pylori* (Genome Sequence of the Lyme Disease Spirochaete, *Borrelia burgdorferi. Nature* 390, 580-586 (1997), The Complete Genome Sequence of the Gastric Pathogen *Helicobacter pylori. Nature* 388, 539-547 (1997)). We have also published the DNA sequence for *Methanococcus jannaschii*, the first archaeal genome to be sequenced, funded by the Department of Energy (DOE); and we will soon be publishing the third DOE-funded genome, *Deinococcus radiodurans* (The Complete Genome Sequence of the Methanogenic Archeon, *Methanacoccus jannaschii. Science* 273, 1058-1073 (1996). No other institution in the world has completed more than one genome.

TIGR has also been funded to sequence human chromosome 16 by the NIH as one of the genome sequencing centers funded through the National Human Genome Research Institute (NHGRI). In support of this effort, DOE has funded TIGR to generate sequence from the ends of 600,000 BACs (bacterial artificial chromosomes) that will form a scaffold linking the human genome sequence together.

PE APPLIED BIOSYSTEMS

During this same timeframe Applied Biosystems had grown as well. The continued expansion of the Human Genome Project, and the use of genomics for research in other areas of biology created huge demand for DNA sequencers. Between 1987 and 1997, more than 6,000 ABI sequencing systems had been sold, giving them the largest installed base of automated sequencers in the world.

In 1993, Perkin-Elmer, a U.S.-based scientific instrument manufacturer, acquired Applied Biosystems and renamed it PE Applied Biosystems. Perkin-Elmer made a significant investment in the life sciences with its acquisition of Applied Biosystems and it has continued to enhance this investment by, for example, investing over $100 million in the last year for research and development to ensure that it continues to develop new, cutting edge technologies. It is one of the these new technologies, the ABI Prism 3700, that will be used for this new venture.

THE HUMAN GENOME PROJECT AND DNA SEQUENCING

As I'm sure you are all familiar, the Human Genome Project has continued to be funded through the DOE and NIH and is now entering its ninth year. This project was officially launched in 1990 as a $3 billion, 15-year federal initiative to map and sequence the complete set of human chromosomes and those of several model organisms. This project was a huge boost to the scientific community and represents a project that, when completed, could have much greater significance to our society than landing on the moon. As a result of this commitment made by the U.S. government, our biotechnology industry, which is holding its annual meeting in New York City this week, leads the world both in the science it undertakes, the jobs it creates, and the products it delivers to improve human health.

Last month a working group completed a review of the draft for the next five-year plan of the Human Genome Project. The program continues to move forward and has made great strides. When it was conceived, very few other organizations, either public or private recognized the value that this activity would have in the scientific and broader communities. Now, largely through the success of this relatively small federal program, whole pharmaceutical companies are restructuring their drug discovery and development process based on genomics.

Unfortunately, when the Human Genome Project was initially explained to the Congress and other organizations a misunderstanding occurred, and the NIH Director, Dr. Harold Varmus, pointed this out at the press briefing we held last month to announce our new venture. The scientists who helped organize this program indicated that sequencing the human genome was the key to improving our knowledge of human biology. This statement has led many to believe that obtaining the complete human DNA sequence would mark the end of the project. In fact, the acquisition of the sequence is only the beginning. The sequence information provides a starting point from which the real research into the thousands of diseases that have a genetic basis can begin. So, the sooner we can get to this starting point, the sooner we can begin to see a payoff in ultimately improving human health.

THE NEW VENTURE AND ITS GOALS

As I earlier indicated, our announcement last month to sequence the human genome within the next three years has been widely reported in both the scientific and popular press. Like the federally-funded project, it captures the imagination. Like the federally-funded project, our goal is not to obtain the sequence for its own sake, but to obtain it to serve as a foundation of data upon which new research into human health can be built. The goal is to develop the definitive resource of genomic and associated medical information that will be used by scientists, in both the public and private sectors, to develop a better understanding of the biological processes in humans and to deliver improved health care in the future.

In addition, this new company intends to build the scientific expertise and informatics tools necessary to extract valuable biological knowledge from this data. This will include discovering new genes, developing polymorphism assay systems, and developing a variety of databases.

There is value in obtaining the sequence of the human genome as quickly as possible--not for the sequences themselves, but for the new research opportunities it will create. There is a significant infrastructure already in place in public sector research institutions that will greatly benefit from this data. Meanwhile, the pharmaceutical and biotechnology industries recognize that the human genome will be the significant resource for future drug discovery and development. Most important, we believe that access to this information is valuable because it will ultimately transform the fundamentals of healthcare delivery and medical practice and improve the lives of millions of people.

The development of a new, fully-automated sequencer by Perkin-Elmer, coupled with the whole-genome shotgun strategy will reduce the costs of operating labor and reagents, while it increases the speed with which sequences can be generated. By building on the resources that have already been developed, such as the significant resource funded by the DOE to sequence the ends of BACs, we have a framework for linking the human genome together, the mechanism for verifying the alignments of sequences on individual chromosomes and internal controls for ensuring the quality of the information that this venture will generate.

The aim of our project is to produce a highly accurate, ordered sequence that spans more than 99.9% of the human genome. The accuracy of this sequence will be comparable to the standard now used in the genome sequencing community of fewer than one error in 10,000 base pairs. We look forward to working with other genome centers to ensure that the sequence meets the requirements of the scientific community for accuracy and completeness.

DATA AVAILABILITY AND INTELLECTUAL PROPERTY

A fact that has often been overlooked or questioned in the press accounts of this venture is that an essential feature of the new company's business plan is to provide public availability of the sequence data. A major consequence of the analysis of data generated by this project will be the creation of a comprehensive human genomic database. Because of the importance of this information to the entire biomedical research community, key elements of this database, including primary sequence data, will be made available. In this regard we will work closely with national DNA repositories like the National Center for Biotechnology Information.

It is our plan to release data into the public domain at least every 3 months including the complete human genome sequence at the end of the project. We also anticipate providing a connect fee for online access to these data and many of the informatics tools that researchers can use to interpret them. We will also market the database system to commercial companies engaged in pharmaceutical and biotechnology research.

A concern that has been raised in many publications is how the intellectual property issues associated with generating the entire human genome sequence will be handled. First, let me just say that I have been associated with intellectual property issues related to DNA sequences from the beginning and have great appreciation for the sensitivities of this concept. By making the sequence of the entire human genome available it makes it virtually impossible for any single organization to own its entire intellectual property. It eliminate the entire speculative nature that is currently associated with patenting DNA sequence information and requires that researchers understand the biology of a sequence before they file a patent application.

Our actions will make the human genome unpatentable. We expect that this primary data will be used by us and others as a starting point for additional biological studies that could identify and define new pharmaceutical and diagnostic targets. Once we have fully characterized important structures (including, for example, defining biological function), we expect to seek patent protection as appropriate. Given the complexity and scope of the information found in the human genome sequence, we expect our efforts to be focused on 100 to 300 targets from among the thousands of potential targets.

CAN THE HUMAN GENOME BE SEQUENCED IN 3 YEARS?

Another question that I have been asked frequently is, can the whole-genome shotgun strategy even work with a genome the size of the human genome? It is our hypothesis that this approach will be successful. In fact, we plan test the effectiveness of this strategy by collaborating with Gerald Rubin of the Howard Hughes Medical Institute and the University of California at Berkeley and the Berkeley *Drosophila* Genome Project to sequence *Drosophila,* another large and complex genome, while we establish the infrastructure for the larger human effort. In addition, this genome will provide us significant insights into the biology of another model organism.

IMPACT ON THE FEDERALLY-FUNDED HGP

Finally, there is the concern that has brought us before you today. How will this new private venture impact the federally-funded Human Genome Project? It is our sincere hope that this program complements the broader scientific efforts to define and understand the information contained in our genome. We recognize that our effort would not even be possible if not for the efforts of those in academia and government who conceived and initiated the Human Genome Project. In fact, the knowledge gained from this effort will provide the key to deciphering the genetic contribution to thousands of human conditions and substantiates and underscores the need to increase the government investment in further understanding of the human genome.

I have heard from different sources that our new venture indicates that the federally-funded program has been a waste of money. I cannot state emphatically enough that our announcement should not be the basis for this claim. Let me explain this by way of an example. Recently, the genome of yeast, *S. cerevisiae,* was completed. This genome was begun before the whole genome shotgun strategy was developed and as a result it took many years to complete. Literally thousands of scientists worked on this project. Does the fact that a faster way to obtain the sequence of the organism they were working on render their work meaningless? Likewise, this new technology and strategy we have announced would have allowed us to sequence the first genome, *H. influenzae*, much more quickly. This fact does not diminish the importance of obtaining the sequence of this organism.

By increasing the speed with which the sequence of the human genome will be obtained, we have not brought any program to completion. We have only helped get everyone to the starting line a little bit sooner. The real race is the one that confronts us each and every day, and that is the one to develop treatments that will help end human suffering brought on by the thousands of diseases that plague humanity.

The impact that our new venture will have on the federally-funded Human Genome Project should be to re-orient it sooner to move beyond DNA sequencing into the research that will help us better understand and treat these diseases.

It is not appropriate to judge the relevance of the Human Genome Project on the basis of our announcement in a retrospective fashion. Without the past we could not be here today. However, it is appropriate to judge the program's relevance in light of our announcement, and others that may come, by the its ability to adapt and work with new initiatives rather than compete against them.

In effect, this new venture is the private sector recognition of the importance of the Human Genome Project. By working closely together, NIH, DOE and other public and private institutions can help meet the goal of having a complete map and sequence of the human genome sooner than anyone ever imagined.

There are many other issues that completing the sequence of the human genome, as well as other genomes, will raise in the very near future. This increased knowledge of evolution, and ultimately ourselves, will likely prompt many questions that society has never even considered. If anything, this new information will require us to strengthen our scientific infrastructure and improve scientific education. We must work to ensure that the science is of the highest quality, appropriately interpreted and peer reviewed. If these areas are addressed, I believe we can appropriately assimilate the wealth of new knowledge and technology that genomics will provide.

CONCLUSION

As I said at the outset, I see the announcement of this new venture as one for which everyone can be proud. It includes the federal government taking the initiative to begin a significant program which is then made more successful by individual creativity and ingenuity, and ultimately is validated by support from the private sector. I hope that after this hearing you view both our announcement and the federal program for which you are responsible as not an "either/or" proposition, but instead focus on how these two activities working in tandem can ultimately improve our lives and those of the generations to come. Thank you.

J. Craig Venter, Ph.D.
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3500

J. Craig Venter, Ph.D., is the Founder, President and Director of The Institute for Genomic Research (TIGR), a not-for-profit, tax exempt basic research institute in Rockville, Maryland. Between 1984 and the formation of TIGR in 1992, Dr. Venter was a Section Chief, and a Lab Chief, in the National Institute of Neurological Disorders and Stroke at the National Institutes of Health (NIH). In 1990, Dr. Venter developed a new strategy for gene discovery. This called *expressed sequence tags* (ESTs) and has revolutionized the biological sciences. Over 72% of all accessions in the public database GenBank are ESTs from a wide range of species including humans, plants and microbes. Using the EST method Dr. Venter and the scientists at TIGR have discovered and published over one half of all human genes. Out of new algorithms developed to deal with 100,000's of sequences TIGR developed the whole genome shotgun method that led to TIGR completing the first three genomes in history.

Dr. Venter recently announced that he signed a letter of intent with Perkin-Elmer for the formation of a new genomics company. The strategy of this company will be centered on a plan to substantially complete the sequencing of the human genome in three years.

Dr. Venter has published more than 150 research articles and is currently tied with Dr. Adams of TIGR as the most cited scientist in biology and medicine. Dr. Venter has received numerous awards and honorary degrees for his pioneering work and has been elected a Fellow of the American Association for Microbiology and the AAAS. Dr. Venter received his Ph.D. in Physiology and Pharmacology from the University of California, San Diego in 1975.

Scientific papers published include:

Complementary DNA Sequencing: "Expressed Sequence Tags" and the Human Genome Project. *Science* 252, 1651-1656 (1991)

Potential Virulence Determinants in Terminal Regions of Variola Smallpox Virus Genome. *Nature* 366, 748-751 (1993)

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science* 269, 496-512 (1995)

Initial Assessment of Human Gene Diversity and Expression Patterns Based Upon 52 Million Basepairs of cDNA Sequence. *Nature* 377 suppl., 3-174 (1995)

The Minimal Gene Complement of *Mycoplasma genitalium*. *Science* 270, 397-403 (1995)

Complete Genome Sequence of the Methanogenic Archeon, *Methanococcus jannaschii*. *Science* 273, 1058-1073 (1996)

The Complete Genome Sequence of the Gastric Pathogen *Helicobacter pylori*. *Nature* 388, 539-547 (1997)

The Complete Genome Sequence of the Hyperthermophilic, Sulphate-Reducing Archaeon *Archaeoglobus fulgidus*. *Nature* 390, 364-370 (1997)

Genome Sequence of the Lyme Disease Spitochaete, *Borrelia burgdorferi*. *Nature* 390, 580-586 (1997)

Complete Genome Sequence of *Treponema pallidum*, the Syphilis Spirochete, *Science* (submitted).

5/98

THE INSTITUTE FOR GENOMIC RESEARCH
GRANTS/PROPOSALS FUNDED AND SUBMITTED

Last Revised: 06/15/98

Source/Title	Period of Project	Grant #	Total Amount
FUNDED			
Department of Energy (Adams) "Expressed Sequence Tags from Human Brain for Genome Mapping"	9/15/92- 3/14/93	DE-FG05-92ER61513	214,364
Department of Energy (Fields) "Identification of Genes in Anonymous DNA Sequences"	4/15/93- 4/14/94	DE-FG02-93ER61566	195,585
AmFAR (Moore) "Examination of Cellular - Viral Protein Associations"	9/1/93- 8/25/95		72,074
Johns Hopkins University (Fraser) "SPORE in Gastrointestinal Cancer"	9/30/93- 9/29/95	5 P50 CA62924-02	100,000
National Science Foundation (Fields) "Integration of Molecular Sequences with Specimen and Taxonomic Data in a Publicly-Accessible Database"	7/1/94- 6/30/96	DEB-9400861	250,000
Department of Energy (Venter) "High-Throughput DNA Sequencing and Characterization of Diverse Microbial Genomes" - REVISED No-cost extension through 2/14/98	11/15/94- 2/14/98	DE-FC02-95ER61962.A000	3,382,525
Department of Energy (Venter) "Mycoplasma genitalium" SUPPLEMENT	11/15/94- 11/14/95	DE-FC02-95ER61962.A001	100,000
Department of Energy (Venter) "High-Throughput DNA Sequencing and Characterization of Diverse Microbial Genomes" - Supplement to Year 3 from above	11/15/94- 11/14/97	DE-FC02-95ER61962.A004	1,000,000
Department of Energy (Venter) "Whole Genome Sequencing of Deinococcus radiodurans" SUPPLEMENT to above	5/15/96- 11/14/97	DE-FC02-95ER61962.A003	2,265,475
SmithKline Beecham Pharmaceuticals (Venter) "Alzheimer's Disease Research"	3/21/95- 3/20/96		767,000
DHHS/NIH (Kirkness) "Characterization of a novel GABA-A receptor subunit; pi"	1/31/96- 12/31/00	1 R29 NS34702-01	520,145
DHHS/NIH (Fraser) "Genetic Organization of Treponema Pallidum"	3/1/96- 2/28/98	1 R01 A140390-01	486,391

THE INSTITUTE FOR GENOMIC RESEARCH
GRANTS/PROPOSALS FUNDED AND SUBMITTED

Last Revised: 06/15/98

Source/Title	Period of Project	Grant #	Total Amount
SUNY (Fraser) "Physical Mapping of Schistosoma Mansoni Chromosomes" (Subcontract)	11/1/97- 4/30/98		25,000
Department of Navy (Tomb) "Genetic Regulation in the Aiptasia pallida Symbiosis" REVISED BUDGET	3/1/96- 2/28/99	N00014-96-1-0604	206,000
The G. Harold and Leila Y. Mathers Charitable Foundation (Fraser)	3/15/96- 3/14/97		320,000
DHHS/NIH (Adams) "Sequencing of Chromosome 16p"04/11/96-03/31/99 REVISED BUDGET Incl. Supplement	4/11/96- 3/31/99	1 R01 HG01464-01	20,165,899
DHHS/NIH (Adams) "Sequencing of Chromosome 16p Supplement	4/01/98- 6/30/98-	3 R01 HG01464-02S1	1,444,305
DHHS/NIH (Adams) "Sequencing of Chromosome 16p Supplement	6/10/98- 6/30/98-	3 R01 HG01464-02S2	452
DHHS/NIH (Fleischmann) "Complete Genome Sequence of Mycobacterium tuberculosis" REVISED BUDGET inc. Malaria supp.	6/15/96- 5/31/99	5 R01 A140125-02	3,267,763
DHHS/NIH (Fleischmann) "Complete Genome Sequence of Mycobacterium tuberculosis" (Ann DeGroot Supplement)	6/1/97- 5/31/98	6 R01 AI 40125-02	142,755
NSF (Venter) "Arabidopsis Genome Sequencing Using Random Shotgun Sequencing of BAC Clones"-REVISED	09/01/96- 8/30/99	DBI-9632085	3,889,391
DOE (Venter) "Arabidopsis Genome Sequencing Using Random Shotgun Sequencing of BAC Clones"-REVISED	09/01/96- 8/31/98	DE-FG02-96ER20249.A001	2,352,600
Department of Energy Subcontract (Adams) "Construction of a genome-wide, highly characterized clone resource for genome sequencing." REVISED BUDGET	9/15/96- 9/14/97	contract#365469	899,945
DHHS/NIH (Clayton) "Whole Genome Sequencing of Vibrio cholerae" - REVISED	12/1/96- 11/30/98	1 R01 A140535-01	1,258,589
DHHS/NIH - (Ketchum) "Complete Genome Sequence of *Enterococcus faecalis* " CFDA 93.856	5/1/97 4/30/99	1 R01 A140963-01	1,181,259

THE INSTITUTE FOR GENOMIC RESEARCH
GRANTS/PROPOSALS FUNDED AND SUBMITTED

Last Revised: 09/15/98

Source/Title	Period of Project	Grant #	Total Amount
Merck Genome Research Institute (Dougherty) " Genome Analysis of *Streptococcus pneumoniae*"	12/01/97- 11/30/98	MGRI Proposal #72	150,000
DHHS/NIH (Quackenbush) "Molecular Analysis of the Human Genome" No-cost extension thru 8/31/98	1/13/97- 8/31/98	7 K01 HG00007-08	75,818
DHHS/NIH (Quackenbush) "MAGE: Molecular Analysis of Gene Expression" 09/30/97-09/29/00	9/30/97- 9/29/99	1 R01 CA77049-01	1,318,968
DHHS/NIH (Lee) "Study of m4 muscarinic receptor mRNA stability" CFDA 93.854	5/1/97- 4/30/02	1-R29 NS35231-01A1	576,057
DHHS/NIH (Lee, N.) "Generation of a Rat EST (REST) Catalog & Rat Gene Index" 09/30/97-09/29/99	9/30/97- 8/31/99	1 R01 HL59781-01	1,026,094
DOE (Ketchum) "Tandem-pore outward-rectifying K+ channels: Molecular partners of the proton ATPase in membrane potential regulation" REVISED BUDGET	5/15/97- 5/14/00	DE-FG02-97ER20265	285,000
Burroughs Wellcome Fund (Gardner) "Complete nucleotide sequence of *Plasmodium falciparum* chromosome14"	7/15/97- 7/14/98		1,250,000
NSF (Ketchum) "Tandem-pore Potassium Channels: Regulators of Electrical and Ionic Homeostasis in Archaea and Eucarya"	8/1/97- 7/31/99	MCB-9723579	180,000
DOE (Venter) "Complete Genome Sequencing of Shewanella putrefaciens" 06/01/97-05/31/99	8/15/97- 8/14/99	DE-FG02-97ER62444	995,480
DHHS/NIH (Tomb) "Porphyromonas gingivalis Genome Project" 07/01/97-06/30/99 RESUBMITTED	9/1/97- 8/31/99	1 R01 DE12082-01A1	1,291,451
DHHS/NIH (Fleischmann) "Complete sequencing and gene expression in M. avium"	9/1/97- 8/31/00	1R01AI41943-01	2,062,805
Department of Energy (Adams) "Construction of a genome-wide, highly characterized clone resource for genome sequencing."	9/15/97- 11/14/98	DE-FC02-97ER62500	2,868,249

THE INSTITUTE FOR GENOMIC RESEARCH
GRANTS/PROPOSALS FUNDED AND SUBMITTED

Last Revised: 06/15/98

Source/Title	Period of Project	Grant #	Total Amount
Department of Defense (Venter) "Malaria Genome Sequencing Project"	12/17/97- 12/16/02	ERMS#972227003	5,125,000
NIH - (Salzberg) (Transfer) "Computational Techniques for Genomic Analysis"	01/01/98- 6/30/99	7 K01 HG00022-04	209,053
Merck Genome Research Institute (Gill) " Genome Analysis of *Staphylococcus aureus*" (Revised Budget)	03/01/98- 2/29/00	MGRI Proposal #73	113,000
NIH-(Gill) " Genome Analysis of *Staphylococcus aureus*" (Revised Budget dated 03/09/98)	04/01/98- 3/31/00	1 R01 AI43567-01	1,305,547
Merck Genome Research Institute (Smith) "Computational Analysis of Intergenic Regions in Microbial Genomes" (Revised Budget)	03/01/98- 2/28/99	MGRI Proposal #74	119,643
DOE(Venter) "An Integrated Program in Microbial Genome Sequencing and Analysis"	02/15/98- 2/14/01		13,500,000
DHHS/NIH (Adams) "African Trypanosome Genome Sequencing"	04/01/98- 3/31/01	1 R01 AI43062-01	2,305,637
DHHS/NIH (Cummings) "Sequence Analysis of Plasmodim falciparum chromosomes 9 and 10"	04/01/98- 3/31/01	1R01A142243-01	2,354,834
DHHS/NIH (Fraser) "Genome Analysis of Chlamydial Species" Changed assignment from 1 R01 HG01784-01	05/01/98- 4/30/00	1 R01 AI43359-01	796,711
Total Funded Grants to Date			62,439,155

THE INSTITUTE FOR GENOMIC RESEARCH
GRANTS/PROPOSALS FUNDED AND SUBMITTED

Last Revised: 06/15/98

Source/Title	Period of Project	Grant #	Total Amount
SUBMITTED/PENDING			
Approved and Pending Award Notice			
DHHS/NIH (Fraser) "Genetic Organization of *Treponema denticola* " 12/1/97-11/30/00	12/1/97- 11/30/00		1,022,956
DOE - (Venter) "Development of Comprehensive Microbial Resource"	02/01/98- 1/31/01		1,560,800

Chairman CALVERT. Dr. Galas.

TESTIMONY OF DAVID J. GALAS, PRESIDENT AND CHIEF SCIENTIFIC OFFICER, CHIROSCIENCE R&D INC., BOTHELL, WA

Mr. GALAS. Mr. Chairman and Mr. Roemer, I certainly welcome the opportunity to testify before the Committee concerning the future of a project so central to the future of, not only the biological sciences but the biotechnology and health care industries of the United States, and it is a pleasure to be here with such a distinguished group.

This is, as is evident, a critical time for this historic project and the attention of Congress, the private sector, and the public sector, and all of the scientific community, is certainly called for to ensure that we make the most of our opportunity here, the opportunity to advance the scientific foundations of these areas that are so important to the health nation.

Now having worked in academia, as well as the private sector, I have witnessed firsthand the effect it has already had on research in the public and private sectors and several of the previous witnesses have cited these. It's become a cliché to call these effects revolutionary and I'm not going to add to any of these clichés, but let me just point out that in this case, almost all of these clichés have been quite accurate.

So why is the Human Genome Project so important and when one summarize this, what is this revolution about? Well, I'd say it's simply about scientists, wherever they are in the life sciences, having the fundamental data close at hand about the information in the human genomes, the genes and regulatory elements, so that they can enable their research into fundamental disease mechanisms, diagnostics, therapeutics, and other fundamental biological mechanisms to an extent never seen before.

Now this genetic information is particularly important to the private sector which is devoted to discovering and developing new therapeutic drugs, among other things. A great deal of money and time is now spent in publicly-supported laboratories and in private companies across the world acquiring genomic information, genomic sequence information piecemeal as it is needed. For example, the availability of the full sequence of the human genome, even a rough version thereof, this past year would have saved our small biotechnology company about, I estimate, about $1.5 million in direct costs and countless months of time on each of several projects. Our work in discovering therapeutics for autoimmune disease, osteoporosis, and other diseases is still a small corner of the biomedical research spectrum, and so these costs to us need to be multiplied by the relative size and number of all involved biotechnology and pharmaceutical companies in this country to see what the direct cost impact on biomedical research would be. Now the indirect costs are also great, as will be the impact on publicly funded research of all kinds. It all adds up to a very large potential savings and some very rough calculations that I made suggest that, perhaps, a year advance in the availability of this information, say in the next year, for purposes of argument, would probably save something like $2 billion in funding in the private sector and, I think, that's quite a conservative estimate.

So the discovery of therapeutics, of course, is not only about money. The savings that arise from better, more effective therapies, and diagnostics that come sooner to the public, and I emphasize the word sooner, must also be a major consideration. The need for widely-available public data resource containing the full complement of human sequence information has never been greater. The announcement by Dr. Venter and his colleagues that they are forming this new enterprise to generate vase amounts of human sequence brings us here today and this project, I'd like to just make a few comments on. This is a most ambitious project, of course, requiring a large number of new things, new automated machines, new computational methods, new significant data production organization, but a relatively small group. It's a difficult undertaking, but as you see, and as you have responded, it is galvanizing, a galvanizing prospect to the entire community.

Now while I cannot directly assess the new technical advances that are cited in their announcement, to me the claims are quite credible and most welcome. And judging from my familiarity with the field, are probably within reach. The scientists involved are experienced, serious, and careful and the prospect of doing what is planned is certainly within what I view as technically feasible and certainly not fanciful. While there will always be debates about how new approaches will work and about the technical details, and these will change, there's no question, from month to month as we go forward, I would say in summary that their proposal seems to be well-founded and plausible.

Now, obviously, the first judgment on their success or failure is going to depend on, on their resolve, their resource commitment, and, finally, on awaiting real results, but it seems to me they have an excellent chance of succeeding and achieving their most important goals. So it is notable and very welcome in addition that the community effort is going to be treated to the availability of the vast amounts of this information as the project goes forward, according to their announcement.

In reaction to that I'd say it's essential that the community and the leadership of the genome project take these prospects very seriously and work both to reform or restrategize about the human genome project strategy, anticipating access to this new data, and to forge close links to the private sector, both sentiments have already been described by the leadership of the project.

So let me just say in emphasis, I do not believe that it is sensible, however, for the federally-supported program either to continue absolutely unchanged with the strategy currently in effect, nor to reduce the level of their efforts. Both of those are very important and I think it's clear from the response so far that at least this general view is shared by both the DOE and the NIH. It seems that the prospect of the private sector sequencing effort has served as quite a useful stimulus to refocusing the Federal effort or at least having a look at the strategy. And I'm sure Dr. Olson will comment on some of these. In my view the, changing the strategy slightly will be very effective and now let me explain what I mean by that in very, in just a few, a few words.

Initially, what's most important in the genome is the location and structure of the functional components, the genes and the control elements. Next most important is the variations that occur in these components, in these component parts, and how they occur in the human population, and the fundamental biological effect on the, on individuals that carry those variations.

Now it is going to be the research, the research work of many decades to understand the basic biological and health effects of these variations. But in achieving the initial goal, the first of these, getting the fundamental understanding information about the genes and their and their control elements, I would argue that it should be the first new goal of the human genome project to focus its attention on getting the first characterization of the genome sequence as quickly as possible. It's been characterized as a first draft, that may be considered to be a pejorative, but I think what we really need is to get that information out as soon as possible and I think plans are under way that could well put this together.

Now reaching this goal in conjunction with the private effort would enable the human genome project to succeed more rapidly than ever, but I think even without that, it's the right thing to do, to reorient towards getting a rapid release of something that some, some call a first draft or an intermediate draft. So this strategy, I think, makes a great deal of sense and let me just summarize the arguments that I'm putting forward for that.

No. 1 is speed. Speed is absolutely critical to the private sector and the public sector. The second one is that it is a major benefit, every piece of new information is a major benefit to the biomedical research community. Third, an effective and positive response to the private sector proposal is also gained by adopting this sort of a strategy. And, finally, future technical effectiveness, I think there are many technical aspects of the revised strategy that stand to provide significant advantages for future sequencing effort once the details were worked through as they will be in the next few years.

Reaching the first goal, however, should be seamless with a follow-on effort to completely fill in the sequence draft, if you will, by producing a very accurate, high quality, and complete reference sequence of the genome. This final project of the human genome program will then become the single most important database of human biology, the complete sequence of our genetic heritage.

Rather than being redundant, the federal program is more relevant than ever, since federal support should now be able to achieve more per dollar spent, and produce a project quite different from what can be expected from the private effort, if the private effort succeeds. I would suggest that more resources should be devoted to the sequencing effort now because the project offers returns soon and the impact of early acquisition of the information will be well worth it.

The prospect before us of a highly-cooperative effort between public and private sectors is one that I think we should seize enthusiastically. Now the federal program appears to be already responding with renewed resolve to this opportunity by rethinking the strategies and there's been a lot of effort, I know, expended on discussing plans for sequencing programs. I applaud this resolve and I expect the genome community at large, both public and private will recognize the critical nature of this moment and seize the opportunity to make the most of it.

This completes my prepared remarks and I'd be happy to answer any questions.

[The prepared statement and attachments of Mr. Galas follow:]

STATEMENT OF

Dr. David J. Galas

President and Chief Scientific Officer
Chiroscience R& D Inc.

1631 220th Street SE
Bothell Washington, 98021

BEFORE THE

COMMITTEE ON SCIENCE

SUBCOMMITTEE ON ENERGY AND ENVIRONMENT

UNITED STATES HOUSE OF REPRESENTATIVES

17 JUNE 1998

Galas, 10 June 1998

Mr. Chairman and Members of the Committee:

I welcome the opportunity to testify before the committee concerning the future of a project so central to the future of medicine, the biological sciences and the biotechnology and health care industries of the United States, the Human Genome Project ("HGP"). This is a critical time in the progress of this historic project and the attention of the congress, the private sector and all the scientific community is called for to insure that we make the most of this opportunity to advance the fundamental scientific foundations of these areas so important to the health of our nation.

I will present here my views on the strategic issues confronting the broader community directly concerned with the project and explain why the impact on the public and private sectors will be so fundamental. I am the President and Chief Scientific Officer of a small biotechnology company in Seattle, Washington. Having worked in academia, as well as the private sector, I have participated in the revolutionary changes in the biomedical sciences engendered by the explosive accumulation of genetic data and of DNA sequence information, and have witnessed, first hand, the effect it has already had on the conduct of research in the public and private sectors. It has become almost a cliché to call these effects revolutionary, but in this case the cliché is accurate. I have served in government, and I am proud to have been in the position of responsibility in DOE now occupied by Dr. Patrinos at the official launch of the Human Genome Project in 1990 by DOE and NIH.

Why is the HGP so important and what is this revolution about? It is simply about scientists and researchers having close at hand the fundamental data about the layout and information content of all the human genome, genes and regulatory elements. This enables research into fundamental disease mechanisms, diagnostics and therapeutics to an extent never seen before. Therefore, this genetic information is particularly important to the private sector devoted to discovering and developing new therapeutic drugs. A great deal of money and time is now spent in private companies across the world acquiring genomic information piecemeal, as it is needed. For example, the availability of the full sequence of the human genome this past year would have saved our small biotechnology company $1.5 million alone in research costs directly expended on sequencing new regions of the genome and countless months of time on each of several projects. Our work towards discovering therapeutics for autoimmune disease, osteoporosis and other diseases is still a small corner of the biomedical research spectrum. These costs to us need to be multiplied by the relative size and number of all the involved biotechnology and pharmaceutical companies in this country to see the direct cost impact on biomedical research - the indirect effects will also be numerous and impressive. It adds up to a very large potential savings, and all of these needs will continue to increase as research advances. In addition, the biomedical research funded by the federal government will also be enabled and accelerated by this information. Therefore, the cost savings to the public and private sectors, in time and money alone, will be enormous. However, the discovery of new therapeutics is not only about money. Savings of another kind, that which arises from better, more effective therapies and diagnostics coming sooner to the public, must also be a major consideration. The need for a widely available, public data resource containing the full complement of human sequence information has never been greater.

What brings us here today is the announcement by Dr. Venter and his colleagues (PE-TIGR) that they are forming a new enterprise to generate vast amounts of sequence data on the human genome in a few short years. This is a most ambitious project, requiring a large number of new automated machines, new computational methods, a significant data production organization and new infrastructure. It is a galvanizing prospect to the entire community. While I am not in a position directly to assess the new technical advances that are cited in their announcement, the claims are both credible in detail and most welcome and judging from my familiarity within the field, are probably well within reach. The scientists involved are experienced, serious and careful and the prospect of doing what is planned is certainly within what I view as technically feasible and certainly not fanciful. While there will always be debates about whether and how new approaches will work and about the technical details, their proposal appears to be well founded and plausible. Final judgment on their success or failure will depend on the resolve and resource commitment of the principals and must, of course, await the first real results, but it seems likely to me that they stand a good chance of succeeding in achieving their most important stated goals. It is notable and very welcome to the entire community that the PE-TIGR effort has made commitment to sharing sequence data with the public HGP.

It is essential that the community and the leadership of the genome project take these prospects very seriously and work both to reform the HGP's strategy anticipating access to this new data and to forge close links to the private sector effort. As I will argue below, I do not believe that it is sensible for the federally supported project either to continue unchanged with the strategy currently in effect, or to reduce the level of their efforts. I think it is clear from the response thus far that this general view is shared by the DOE and NIH alike. They appear to be responding with an eminently sensible attempt at revision of the strategy for sequencing and a commitment to take advantage of whatever new sequencing capacity and data release comes from the private effort. It seems that the prospect of the private sector sequencing effort has served as a beneficial stimulus to refocus the federal effort on a strategy that will, in my view, maximize the effectiveness of the project whether or not the private effort reaches their stated goals. If they do reach these goals the strategy will greatly advance the rate of accumulation of useful data and hasten the day of the first completion of the sequence of the human genome.

Initially, what is most important in the genome is the location and structure of the functional components - the genes and their control elements. Next important is the variations that occur in these component parts in the human population and the fundamental biological effects on the individuals that carry these variations. It is these variations that make each of us distinct in our good health and strengths, and our susceptibility to disease and ill health. It will be the research work of many decades to understand the extent and the basic biological and health effects of these variations - this work will be a large part of the future of medical research.

The initial goal of the HGP sequencing effort is to provide the initial blueprint, the basic sequence, not the myriad of sequence variations. While many basic researchers and companies alike, us included, are focused on detecting and understanding consequences of these many small variations in the human genome, called single nucleotide polymorphisms or SNPs, we all need the initial sequence to progress this next wave of biomedical research. Therefore, I argue that it should be the essential primary goal of the HGP to focus its attention on how to arrive at the first initial characterization of the genome sequence as quickly as possible, whether or not the private effort contributes in the long run to reaching this goal. Reaching this goal in conjunction with the private effort, however, would enable the HGP to succeed more rapidly than ever, but even without the impetus of the prospect of the private effort the HGP should be re-oriented to this primary goal - to obtain an initial "first draft" of the human genome as soon as possible. Even a rough "first draft" would be absolutely invaluable to the broad biomedical community. It appears that the prospect that brings us here today has galvanized the HGP into considering a strategy like this in any case and one that could, with public-private cooperation, lead to a much more rapid achievement of this initial goal. This strategy makes sense.

To summarize the arguments for a refocused HGP strategy:

1. Speed. The critical information will be available sooner, probably 95% within 3 years.
2. A major benefit to biomedical research. The benefits of locating genes and control elements sooner will substantially advance all biomedical research sectors.
3. An effective and positive response to the PE-TIGR proposal. Refocus of the HGP strategy takes advantage of the opportunity to leverage the private sector investment into a valuable public resource.
4. Future technical effectiveness. There are many technical arguments for the revised strategy that stand to provide advantages for future sequencing efforts once the details are worked through.

The achievement of the initial goal of a “first draft” should in no way mark the end of the project. It is important that the reaching of the first goal be seamless with a continuing, follow-on effort to complete the sequence “draft” by producing a very accurate, high-quality, complete reference sequence of the genome. Finishing this final product is just as important as the initial goal and will be easier and less expensive than it is now. This final product of the HGP will then become the single most important database of human biology, the complete sequence of our genetic heritage.

Rather than being redundant, the federal HGP is more relevant than ever, since federal support should now be able to achieve more per dollar spent, and produce a product quite different from what can be expected from the private effort. I suggest that the early prospect of completion that arises from the private proposal should be met with increased funding for the federal project, subject to successful completion of the new planning effort that is underway. The changes should not, however, end there. The prospect before us of a strong, highly cooperative effort between the public and private sectors is one that we should seize enthusiastically. Public-private sector cooperation too often is afflicted with bureaucratic viscosity, management difficulties and basic problems in reaching the stated goals. To my view, this opportunity appears to be one that will lend itself well to avoiding these pitfalls. The benefits to both sides and to the public at large, of a successful endeavor are indeed great and the commitments and progress will be visible and accountable in large measure by both sides.

The federal program appears to be already responding with renewed resolve to this opportunity by rethinking the strategy and replanning the sequencing programs and I expect the genome community at large, both public and private, will recognize the critical nature of this moment and seize the opportunity to make the most of it.

This completes my prepared testimony. I would be happy to answer any questions.

David J. Galas, Ph.D.
Biography

David J. Galas, Ph.D. is currently Executive Director of Darwin Discovery, Chiroscience Group plc. and is also President and Chief Scientific Officer at Chiroscience R&D, Inc., formerly Darwin Molecular Corporation. Before joining Darwin in August of 1993, Dr. Galas served as Director for Health and Environmental Research, U.S. Department of Energy, where he had responsibility for the Human Genome Project and all biological and environmental research. He assumed his position with the Department of Energy in 1990 after nine years on the faculty of the University of Southern California, Department of Biological Sciences as Professor of Molecular Biology. Dr. Galas also held positions at the University of California – Lawrence Livermore National Laboratory and the University of Geneva (Switzerland) before joining USC in 1981.

Dr. Galas has been a member of many federal and academic advisory groups including the National Biotechnology Policy Board, the National Cancer Advisory Board, and the National Academy of Science Research Council Board on Biology. He chaired the biotechnology Research Subcommittee of the Federal Coordinating Council on Science and Technology.

His research interests have included the study of the transposition of genetic elements and their consequences, and the study of DNA-protein interactions. He has developed several techniques used in molecular biology research, including the widely used DNA "footprinting" technique, a method often used to define DNA sequence-specific sites for DNA-binding proteins controlling gene expression for DNA replication and recombination. He has a long-term interest in the development of interdisciplinary research in the biological sciences and the applications of diverse technologies to biological problems.

Dr. Galas received his B.A. in Physics at the University of California at Berkeley in 1967 followed by a Master's degree and a Ph.D. in Physics from the University of California, Davis-Livermore in 1968 and 1972 respectively.

CHIROSCIENCE

David J. Galas, Ph.D.
President &
Chief Scientific Officer

June 15, 1998

VIA FACSIMILE
(202) 226-6983

Mr. Ken Calvert
Chairman
Subcommittee on Energy and Environment
Suite4 2320 Rayburn House Office Building
Washington, DC 20515

Dear Mr. Calvert:

In reference to your letter dated June 10, 1998 please be advised that Chiroscience R&D, Inc. has not received federal funding during the current and two preceding fiscal years relating to this testimony.

Sincerely,

David Galas/eac

David J. Galas

Chiroscience R & D, Inc. 1631 220th Street SE, Bothell, WA 98021, USA
Tel. 425 489 8000 Fax. 425 489 8018
A member of the Chiroscience group of companies

Chairman CALVERT. Thank you, Doctor.
Doctor Olson.

TESTIMONY OF MAYNARD V. OLSON, PROFESSOR OF MEDICAL GENETICS AND GENETICS, DEPARTMENT OF MOLECULAR BIOTECHNOLOGY, AND DIRECTOR, GENOME CENTER, UNIVERSITY OF WASHINGTON, SEATTLE, WA

Mr. OLSON. Thank you, Mr. Chairman. I'm here to provide the perspective of an academic researcher who has been involved in what is now called genome analysis for over 20 years. Indeed, my involvement dates to a time when the term genome was rarely used, even in scientific circles, and had yet to have any impact whatsoever on public discourse. Since then, of course, the times have changed as this hearing and the intensive press coverage of the Perkin-Elmer announcement indicate. They've changed, perhaps, foremost because the singular historical opportunity that we now face to unravel the molecular details of how the information is stored and what the information is that guides the transformation of a fertilized egg into a fully-developed human being has caught both the popular and the scientific imagination.

More practically, and, perhaps, more forcefully in the short run, times have changed as the immediate value of the data produced by genome analysis has become evident, particularly the value of DNA sequence data. These data have a high scientific value and also a high value in dollars, yen, and Euros. Thus, entering a major participation of the commercial, injecting a major participation of the commercial sector into what had previously been predominantly a basic science initiative.

Congress now faces a new challenge of understanding and responding to a scientific environment in the human genome project that has all of chaos that comes with scientific and policy success. My basic message in this turbulent environment if quite system and that is that the system is working. It is important to keep in mind that biomedical research in the United States derives its formidable strength from the synergy between three sectors, the biotechnology industry, the more traditional pharmaceutical industry, and academic and publicly-supported research. All of these sectors are scrambling in their own ways to adjust to our sudden ability to produce DNA sequence on a large scale. In this context the Perkin-Elmer announcement is a bold example of the response of the biotech sector to these opportunities.

Perkin-Elmer is adopting here an overtly biotech style of operation despite its roots as a manufacturer of scientific instruments and reagents. It's a hallmark of the biotech style that time is of the essence and publicity is a key tool for influencing events. Those of who are watching this spectacle from the sidelines should certainly wish Perkin-Elmer well. The company's investment will surely lead to faster testing of new reagents and instrumentation and also will produce much data that will be of both commercial value and basic scientific interest.

However, the excitement generated by the well-orchestrated public relations campaign surrounding this announcement should not disguise that what we have at the moment is neither new technology nor even new scientific activity. What we have is a press release. And I believe when I speak for many academic spectators when I say I look forward to a transition from plans to reality. In short, show me the data.

I cannot emphasize too strongly that science by press release and, worse yet, science policy by press release is not a path that the United States Congress or the federal agencies wants to walk down. I believe that the overwhelming risk for the publicly-funded program is one of overreaction. What the Perkin-Elmer initiative offers with the greatest probably is that the immediate needs of the biological community during a period of a few years, roughly in the interval 2000 to 2003 may be better met than would otherwise have been the case. And I hope that the project is successful and that the data are sufficiently accessible to the scientific community that this promise is met.

However, in the larger scheme of the Human Genome Project, we would all be unwise to focus on so transient the contribution. The case for the transience of these data's value lies in one's assessment, in advance, of any real basis to make such a judgment of the likely quality of the final product, as has mentioned repeatedly by others at the table and will be a subject of intensive technical discussion for some years to come.

I, frankly, am a skeptic that the approaches as publicly described will lead to a product of sufficient quality to meet the long-term needs of the scientific community. I'm prepared to be proven wrong, as any scientist must be, but I am comfortable predicting that this approach, as the downside of its efficiency, will encounter reasonably catastrophic problems at the stage of which the tens of millions of independent sequencing tracks need to be melded together to produce a composite view of the human genome.

To be specific, I'm comfortable predicting that there will be over 100,000 serious gaps in the final product and in this context, I define a serious gap as one in which there is uncertainty even as how one should orient and align the islands of assembled sequence between the gaps. Furthermore, I'll predict that a substantial fraction, particularly the smaller islands of sequence of produce will be misassembled, that is they will not actually correspond to the organization of the human genome and I say these things being thoroughly familiar, and admiring, TIGR's success in sequencing bacterial genomes by what superficially would appear to be a similar strategy.

I want to emphasize that even such data will certainly have considerable biological utility and it may prove to be a major help in the final push toward a high quality human sequence, although I would also emphasize that this prospect is somewhat less certain. Experience has tended to show that large amounts of low-quality sequence data are a poor substitute for smaller amounts of high-quality data collected for the specific purpose of assembling a contiguous, accurate sequence which I believe should continue to be, with a minimum of distractions, the focus of the publicly-funded effort.

Clearly, as time develops, if data from this private initiative proves to be of clear utility in achieving that publicly-financed goal, other strategies should, and will, adapt. I want to emphasize that there are two reasons to aim high in terms of the quality of the final human sequence. And, frankly, I am much more concerned about the force of these arguments than I am about the opportunity costs, although I acknowledge there will be opportunity costs, associated with relatively transient delays in the availability of the final product.

The two reasons have to do, first, with deferred costs as a practical reason. A human sequence that has many deficiencies will defer for decades to come, throughout the biomedical research enterprise, the need to fix small problems as they are encountered by individual investigators. The other argument is perhaps even more important in taking the broad view of public policy in this matter. And that is that all of us, as we build the total package of activity in the public sector, the private sector, throughout biomedical and agricultural product research, we need, collectively to achieve an extremely high standard in human genetics. We should start with an extremely high scientific standard and not waver in our commitment to that goal.

The human genome sequence is part of that commitment. A more important part, built upon it, will be our study of human variation and the biological consequences of that variation.

So, I have some additional comments in my written records, but I hope, for the purposes of this hearing, that the Congressional message to the federal agencies responsible in this area will be that you are proud of your institution's role in initiating this project and look forward, as I do, to the production of a sequence that is freely accessible to all sciences, delivered on schedule, and of impeccable quality.

Thank you.

[The prepared statement and attachments of Mr. Olson follow:]

Testimony of Maynard V. Olson before the House Committee on Science, Subcommittee on Energy and Environment, scheduled for June 17, 1998

The Human Genome Project has come a long ways since its fragile beginnings a decade ago. In its early years, the proposal to develop a complete DNA sequence of the human genetic material often seemed an idea ahead of its time: the project's feasibility could reasonably be questioned, there was little support amongst rank-and-file biologists, and the pharmaceutical and agricultural-products industries were disengaged. Now, residual technical arguments involve minor squabbles between experts, basic and applied biological research is reorganizing itself around the assumption that complete genome sequences will soon be available for all intensively studied organisms, and the commercial sector has emerged as a major player in large-scale genome analysis. Indeed, we not only now have a vigorous biotech industry—in which the United States is the undisputed world leader—but a whole tier of "genomics" companies created to meet the insatiable demand for specialized data about genomes that has arisen throughout the biotechnology, pharmaceutical and agricultural-products industries.

It is worth reflecting briefly on the reasons for this success. First, there are the scientific fundamentals. We have only known for a few decades that all life is based on digital information—the "base-four" code of DNA sequence that is now featured even on movie marquees (as in the movie title "GATTACA," which is simply a short bit of DNA sequence expressed with the four standard symbols G, A, T, and C). The information present in a human sperm or egg cell is encoded in 3 billion G's, A's, T's, and C's. Thus, the total information content of the human genome is only 750 Megabytes—about the capacity of a compact disc—an awe-inspiring level of data compression.

Although the challenge of interpreting the human sequence will remain a central preoccupation of science for centuries to come, available sequence data already yield rich dividends. Most profoundly, computer-based methods of sequence comparison frequently allow detection of functionally informative similarities between genes discovered in different organisms. This feature of DNA sequences allows biologists studying human diseases to infer important lessons about the molecular basis of these pathological processes through gene-to-gene comparisons with the richly informative data already available about the genes of "model" organisms such as yeast and fruit flies.

A former member of this institution, Rep. Claude Pepper, deserves great credit for having recognized that biological research needed to be led aggressively into the information age. His support for establishment of the National Center for Biotechnology Information at the National Library of Medicine is one of the great success stories of proactive involvement by the Congress in the building of research infrastructure. The Wold Wide Web site of the NCBI, on which DNA-sequence comparisons are the central activity, has become a major epicenter of biological research.

As the NCBI story illustrates, the present success of genome analysis has roots in policy as well as science. In the Human Genome Project, Congress was actually ahead of the majority of scientists in recognizing that it was time to move boldly to create an information-based future for biomedical research. The establishment of the Human Genome Project, which led in a few years to the creation one of the NIH's most dynamic and forward-looking Institutes, the National Human Genome Research Institute, was the work of a relatively small group of committed scientists and federal officials, who brought a strong case to Congress and received an equally strong response. This response was all the more impressive given the draconian budgetary constraints that had to be overcome to bring the Human Genome Project into existence.

The Congress now faces the new challenge of understanding and responding to a scientific environment in the Human Genome Project that has all the roiling chaos that comes with scientific and policy success. My basic message in this turbulant environment is quite simple: the system is working.

Biomedical research in the United States derives its formidable strength from the synergy between three dynamic sectors: academic research, the biotechnology industry, and the pharmaceutical industry. Academic research, with its reliance on federal funding and the stewardship of a highly evolved resource-allocation system administered by the NIH and other federal agencies, is clearly "the goose that laid the golden egg." The pharmaceutical industry provides a powerful engine for translating new research into safe-and-effective products. As the pace of biological research has accelerated following the development of recombinant-DNA techniques and the introduction of other new research tools, a whole industry—the increasingly important biotech sector—has arisen to respond rapidly to new commercial opportunities. This sector is characteristically quicker on its feet and more willing to take large business risks than the pharmaceutical industry. Time will tell whether the pharmaceutical and biotech sectors ultimately merge or retain their currently distinct identities.

The present landscape in the Human Genome Project illustrates well the operation of all three sectors. The academic sector is focused on the creation of a high-quality reference sequence of the human genome, presently targeted for completion in 2005. This still-ambitious goal is defined in terms of rigorous quality-control standards enforced through a vigorous process of peer-reviewed scientific performance and peer-assessment of data quality. The academic sector is also responsible for the critical task of training a growing cohort of young scientists who can lead genome analysis into its open-ended future. Similarly, academic research is the incubator in which new technical approaches and new applications of genome analysis to biology are under development.

Increasingly, the pharmaceutical industry is redirecting long-term drug-discovery programs to exploit the new opportunities provided by an avalanche of sequence data, data that are leading daily to the discovery of new genes, new proteins, and new functional dimensions to life processes. In addition to its primary reliance on public-domain sequence data, the pharmaceutical industry is building in-house data-collection capabilities—and even more dramatically—pursuing such data through a host of contracts, partnerships, and other relationships with biotech and genomics companies.

It is against this background that the recently announced Perkin Elmer initiative to accumulate a large database of DNA sequences sampled directly from the human genome should be viewed. Although traditionally a manufacturer of scientific instruments and research reagents, Perkin Elmer is adopting, in this venture, an overtly "biotech" style of operation. The business risks are considerable since it remains unclear how the company will recover its substantial investment. Furthermore, as is a hallmark of biotech research, time is of the essence and publicity is a key tool for influencing events. Those of us who are watching this spectacle from the sidelines (i.e., as neither participants nor competitors) should wish Perkin Elmer well. The company's investment will surely stimulate rapid reduction-to-practice of new reagents and instrumentation and will also produce much data that will be both of commercial value and basic scientific interest. However, the excitement generated by the well-orchestrated public-relations campaign surrounding the Perkin Elmer announcement should not disguise that what we have at the moment is neither new technology nor even new scientific activity: what we have is a press release. I believe that I speak for many academic spectators when I say that I look forward to a transition from plans to reality. In short, "Show me the data."

The risk here for the publicly funded program is one of overreaction. What the Perkin Elmer initiative offers is the possibility that the immediate needs of the biological community during a period of 2-3 years, roughly in the interval 2000-2003, may be better met than would otherwise have been the case. I hope that the project is successful and that the data are sufficiently accessible to the scientific community that this promise is met. However, in the larger scheme of the Human Genome Project, we would all be unwise to focus on so transient a contribution.

The case for the transience of these data's value lies in the likelihood that they will be of poor quality. While I am prepared to be proven wrong, as any scientist must be, I am equally prepared to put my reputation as a scientific prognosticator on the line in predicting that the Perkin Elmer initiative will fail to produce a sequence of the human genome that will meet the long-term needs of the scientific community. Specifically, I predict that the proposed technical strategy for sampling human DNA sequences will encounter catastrophic problems at the stage at which the tens of millions of individual tracts of DNA sequence must be assembled into a composite view of the human genome. Based on extensive experience with the assembly of composite human DNA sequences in our genome center and other laboratories, I predict that there will be over 100,000 "serious" gaps in the assembled sequence: a "serious" gap, in this context, is one in which there is uncertainty even as to how to orient and align the islands of assembled sequence between the gaps. Furthermore, I predict that a significant fraction of the small islands between serious gaps will be misassembled (i.e., they will not actually correspond to the organization of the human genome).

Even such fragmentary data will certainly have considerable biological utility. Furthermore, it may prove to be a substantial help in the final push toward a high-quality human sequence, although this prospect is less certain. Experience has tended to show that large amounts of low-quality sequence data are a poor substitute for smaller amounts of high-quality data collected for the specific purpose of assembling a contiguous, accurate sequence.

It is of the utmost importance that a vigorous public effort be maintained that is directed toward the development of a sequence that will meet the test of time. There are two compelling rationales for aiming high in terms of the quality of this sequence. In practical terms, any other approach will defer large costs, diffusing them across the biomedical research enterprise for decades to come as individual investigators are left to complete and correct the reference sequence in regions of the genome where the data are inadequate to meet their particular needs. Perhaps still more important is the need to set a high standard in all aspects of human genetics, starting with an unwavering commitment to quality in the Human Genome Project's flagship mission. Although I have confidence that the spectacular advances we are currently witnessing in human genetics will lead to great public benefit, I do not share the view—expressed in some quarters—that the speed of generating data must take precedence over all other considerations. An element of caution in developing this first comprehensive view of the human genetic material is advisable. High scientific standards tend to be infectious. I would like the legacy of my involvement in the Human Genome Project to be a product that will not only facilitate the research of future scientists but will also inspire them to set a similarly high scientific standard as they interpret the sequence and study its variation from one human to another and the effects of that variation on human biology.

For its part in bringing about this future, I would advise Congress to wait and watch rather than to attempt to provide detailed guidance to the involved agencies. At root, many of the issues are deeply technical and Congress is the wrong forum in which to debate the relative merits of capillary-gel electrophoresis vs. slab-gel electrophoresis, whole-genome "shotgun" sampling vs. a clone-by-clone approach, and so forth. The agencies need a more general sense of how Congress views the public benefit associated with the Human Genome Project. I hope that the Congressional message will be that you are proud of your institution's role in initiating this project and look forward, as I do, to the production of a sequence that is freely accessible to all scientists, delivered on schedule, and of impeccable quality.

I would like to close by identifying three areas of concern that I do think bear further scrutiny by appropriate Congressional processes. First, I think there is a strong case for increased funding for the National Human Genome Research Institute, although my argument for increased funding would differ from that of many of my colleagues. I believe that the current NHGRI budget is actually adequate, in combination with funding through other channels, to produce a quality human sequence by 2005. Given the large technical uncertainties, I think the National Research Council Committee on the Mapping and Sequencing of the Human Genome, on which I had the honor of serving, did a good job of projecting the cost of the Human Genome Project. Indeed, it also did a good job of estimating the time required to complete the project. I doubt that the current schedule could be much accelerated without encountering human-resource bottlenecks that would be difficult to overcome.

However, I am concerned that without expanded funding, the peak phase of data production for human sequencing, will drain other valuable activities at the NHGRI. The NRC Committee did not fully envision the rapidity with which genome analysis would open up new opportunities in biological research. Indeed, the Perkin Elmer proposal is but one symptom of the magnitude and immediacy of these opportunities. While moving ahead toward its flagship goal of producing a quality human sequence, the NHGRI also faces increasing responsibilities to identify and stimulate research avenues opened by the early successes of the Human Genome Project. These opportunities include development of new technology, improved computational methods for analyzing DNA sequence, approaches to the comprehensive functional analysis of genomes, and—perhaps most profoundly—characterization of natural variation in human DNA. In my view, the strongest case for increased NHGRI funding lies in its excellent track record and the continuing expansion of research opportunities in areas that go beyond the Institute's core mission but which provide critical links between the emerging human sequence and the rest of biological research.

Two other issues, which are illustrated by, but not narrowly related to, the Perkin Elmer initiative bear Congressional attention. The most important concerns the influence of intellectual-property law on the research enterprise. Particularly in areas where the interests of the three major sectors of biomedical research—academe, the pharmaceutical industry, and the biotechnology industry—diverge, there are increasing signs of trouble. The pharmaceutical industry has legitimate concerns that it has become too easy for biotechnology companies to acquire valuable intellectual-property rights through cream-skimming research investments. Continuation of the current system risks the accumulation of disincentives for drug development in certain areas or, alternately, diversion of the attention of pharmaceutical companies into purely defensive acquisition of its own tenuous intellectual-property claims. Academic research faces other concerns. Foremost amongst these are situations in which the conduct of basic research in the non-profit sector—the very research on which our current success rests—is distorted by conflicts over intellectual property and access to data. In the worst cases, commercial owners of intellectual property are using their property to attempt to impede research in the non-profit sector when they do not see that research as compatible with their short-term interests.

A more direct warning posed by the Perkin Elmer initiative is that academic researchers risk losing equal access to critical research tools. These tools, such as advanced instrumentation for DNA analysis, are increasingly seen as a means through which their developers can acquire intellectual property rather than as products in their own right. Perhaps if the microscope were a contemporary invention, we would find optical companies competing to sell images rather than microscopes. Basic scientists need access to state-of-the-art research tools, not just to the output of these tools. However, the tools themselves are now universally refined, manufactured, and marketed by private companies rather than by basic researchers themselves. Hence, tool-making companies are in a powerful position to influence the directions that basic research takes and the distribution of that research between the non-profit and for-profit sectors. Instrumentation provides one simple illustration of this dynamic; however, even more problematic situations arise in areas such as reagents, analytical processes, and reference databases. There are no simple answers to the resultant dilemmas, but the public interest in keeping basic researchers well equipped to do their work is clear. The United States is the world leader in an area that is central to the human future—biomedical and agricultural research—and it has gained this enviable position by coupling the world's strongest system of research universities to an aggressive commercial sector. Effort expended fine-tuning the relationship between these parties will be effort well spent.

2/98

MAYNARD V. OLSON

Education

B.S.: California Institute of Technology, Pasadena, CA; June, 1965; major field, chemistry

Ph.D.: Stanford University, Stanford, CA; January, 1970; major field, inorganic chemistry; thesis advisor, Henry Taube; thesis title: I. Studies with maleate as a ligand; II. ^{17}O magnetic resonance of aqueous solutions of V(II) and Cr(III).

Awards and Honors

Undergraduate:	Graduated from Caltech with Honors (1965)
Graduate:	National Science Foundation Graduate Fellowship (1965-1969)
Postdoctoral:	National Institutes of Health Individual Postdoctoral Fellowship (1977-1979)
Professional:	Genetics Society of America Medal (1992) Fellow of the American Association for the Advancement of Science (1993) National Academy of Sciences (1994)

Positions Held

9/69 - 1/76	Instructor and Assistant Professor, Department of Chemistry, Dartmouth College, Hanover, New Hampshire
9/74 - 8/75	Visiting Scholar, Department of Genetics, University of Washington, Seattle, Washington
2/76 - 7/79	Research Associate, Department of Genetics, University of Washington, Seattle, Washington
8/79 - 8/92	Assistant Professor, Associate Professor, and Professor of Genetics, Washington University School of Medicine, St. Louis, Missouri
10/89 - 8/92	Investigator, Howard Hughes Medical Institute at Washington University, St. Louis, Missouri
9/92 -9/97	Professor of Molecular Biotechnology, University of Washington, Seattle, Washington
9/92-	Professor of Medicine (Division of Medical Genetics), University of Washington, Seattle, Washington
8/96-	Adjunct Professor of Computer Science, University of Washington, Seattle, Washington
7/97-	Professor of Genetics, University of Washington, Seattle, Washington

Professional Service

(1987 - 1988): National Research Council Committee on Mapping and Sequencing the Human Genome
(1987 - 1988): Genetics Study Section of the National Institutes of Health
(1989 - 1992): Program Advisory Committee on the Human Genome of the National Institutes of Health
(1994-): National Research Council Government-University-Industry Research Roundtable Council
(1994-): Chairman, Genome Research Review Committee of the National Human Genome Research Institute, National Institutes of Health

Society Memberships

American Association for the Advancement of Science
Genetics Society of America

Editorial Boards

Genomics
Genome Research
Human Genetics

Publications (Research Papers)

Olson, M.V., Kanazawa, Y., and Taube, H. (1969). ^{17}O magnetic resonance of aqueous solutions of vanadium(II) and chromium(III). *J. Chem. Phys.* **51:** 289-296.

Olson, M.V. and Taube, H. (1970). Hydration and isomerization of coordinated maleate. *J. Amer. Chem. Soc.* **92:** 3236-3237.

Olson, M.V. and Taube, H. (1970). The chromium(II) reduction of maleatopentaammine-cobalt(III). *Inorg. Chem.* **9**: 2072-2081.

Olson, M.V. (1973). Reaction between ethylenediaminetetraacetic acid and carboxylato-pentaaquochromium(III) complexes. *Inorg. Chem.* **12:** 1416-1423.

Olson, M.V. and Behnke, C.E. (1974). Kinetics of the spontaneous ring-closing and aquation reactions of malonatopentaaquochromium(III). *Inorg. Chem.* **13:** 1329-1334.

Goodman, H.M., Olson, M.V., and Hall, B.D. (1977). Nucleotide sequence of a mutant eukaryotic gene: the yeast tyrosine-inserting ochre suppressor SUP4-o. *Proc. Natl. Acad. Sci. USA* **74:** 5453-5457.

Olson, M.V., Montgomery, D.L., Hopper, A.K., Page, G.S., Horodyski, F., and Hall, B.D. (1977). Molecular characterisation of the tyrosine tRNA genes of yeast. *Nature* **267:** 639-641.

Olson, M.V., Hall, B.D., Cameron, J.R., and Davis, R.W. (1979). Cloning of the yeast tyrosine transfer RNA genes in bacteriophage lambda. *J. Mol. Biol.* **127:** 285-295.

De Robertis, E.M. and Olson, M.V. (1979). Transcription and processing of cloned yeast tyrosine tRNA genes microinjected into frog oocytes. *Nature* **278:** 137-143.

Olson, M.V., Loughney, K., and Hall, B.D. (1979). Identification of the yeast DNA sequences that correspond to specific tyrosine-inserting nonsense suppressor loci. *J. Mol. Biol.* **132:** 387-410.

Olson, M.V., Page, G.S., Sentenac, A., Piper, P.W., Worthington, M., Weiss, R.B., and Hall, B.D. (1981). Only one of two closely related yeast suppressor tRNA genes contains an intervening sequence. *Nature* **291:** 464-469.

Shalit, P., Loughney, K., Olson, M.V., and Hall, B.D. (1981). Physical analysis of the CYC1-sup4 interval in *Saccharomyces cerevisiae. Mol. Cell. Biol.* **1:** 228-236.

Sandmeyer, S.B. and Olson, M.V. (1982). Insertion of a repetitive element at the same position in the 5'-flanking regions of two dissimilar yeast tRNA genes. *Proc. Natl. Acad. Sci. USA* **79:** 7674-7678.

Brodeur, G.M., Sandmeyer, S.B., and Olson, M.V. (1983). Consistent association between *sigma* elements and tRNA genes in yeast. *Proc. Natl. Acad. Sci. USA* **80:** 3292-3296.

Carle, G.F. and Olson, M.V. (1984). Separation of chromosomal DNA molecules from yeast by orthogonal-field-alternation gel electrophoresis. *Nucleic Acids Res.* **12:** 5647-5664.

Fischhoff, D.A., Waterston, R.H., and Olson, M.V. (1984). The yeast cloning vector YEp13 contains a tRNA$_3^{Leu}$ gene that can mutate to an *amber* suppressor. *Gene* **27:** 239-251.

Gray, A.J., Beecher, D.E., and Olson, M.V. (1984). Computer-based image analysis of one-dimensional electrophoretic gels used for the separation of DNA restriction fragments. *Nucleic Acids Res.* **12:** 473-491.

Shaw, K.J., and Olson, M.V. (1984). Effects of altered 5'-flanking sequences on the in vivo expression of a *Saccharomyces cerevisiae* tRNA^{Tyr} gene. *Mol. Cell. Biol.* **4:** 657-65.

Carle, G.F. and Olson, M.V. (1985). An electrophoretic karyotype for yeast. *Proc. Natl. Acad. Sci. USA* **82:** 3756-3760.

Helms, C., Graham, M.Y., Dutchik, J.E., and Olson, M.V. (1985). A new method for purifying lambda DNA from phage lysates. *DNA* **4:** 39-49.

Burke, D.T. and Olson, M.V. (1986). Oligodeoxynucleotide-directed mutagenesis of *Escherichia coli* and yeast by simple cotransformation of the primer and template. *DNA* **5:** 325-332.

Carle, G.F., Frank, M., and Olson, M.V. (1986). Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. *Science* **232:** 65-68.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83:** 7826-7830.

Burke, D.T., Carle, G.F., and Olson, M.V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806-812.

Graham, M.Y., Otani, T., Boime, I., Olson, M.V., Carle, G.F., and Chaplin, D.D. (1987). Cosmid mapping of the human chorionic gonadotropin beta subunit genes by field-inversion gel electro-phoresis. *Nucleic Acids Res.* **15:** 4437-4448.

Johnson, D.I., Jacobs, C.W., Pringle, J.R., Robinson, L.C., Carle, G.F., and Olson, M.V. (1987). Mapping of the *Saccharomyces cerevisiae* CDC3, CDC25, and CDC42 genes to chromosome XII by chromosome blotting and tetrad analysis. *Yeast* **3:** 243-253.

Riles, L. and Olson, M.V. (1988). Nonsense mutations in essential genes of *Saccharomyces cerevisiae*. *Genetics* **118:** 601-607.

Brownstein, B.H., Silverman, G.A., Little, R.D., Burke, D.T., Korsmeyer, S.J., Schlessinger, D., and Olson, M.V. (1989). Isolation of single-copy human genes from a library of yeast artificial-chromosome clones. *Science* **244:** 1348-1351.

Riethman, H.C., Moyzis, R.K., Meyne, J., Burke, D.T., and Olson, M.V. (1989). Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast artificial-chromosome vector. *Proc. Natl. Acad. Sci. USA* **86:** 6240-6244.

Green, E.D. and Olson, M.V. (1990). Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. *Proc. Natl. Acad. Sci. USA* **87:** 1213-1217.

Green, E.D. and Olson, M.V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: A model for human genome mapping. *Science* **250**: 94-98.

Drury, H., Green, P., McCauley, B., Olson, M.V., Politte, D.G. and Thomas, L.J. Jr. (1990). Spatial normalization of one-dimensional electrophoretic gel images. *Genomics* **8**: 119-126.

Imai, T. and Olson, M.V. (1990). Second-generation approach to the construction of yeast artificial-chromosome libraries. *Genomics* **8**: 297-303.

Link, A.J. and Olson, M.V. (1991). Physical map of the *Saccharomyces cerevisiae* genome at 110-kb resolution. *Genetics* **127**: 681-698.

Huxley, C., Hagino, Y., Schlessinger, D. and Olson, M.V. (1991). The human HPRT gene on a yeast artificial chromosome is functional when transferred to mouse cells by cell fusion. *Genomics* **9**: 742-750.

Gnirke, A., Barnes, T.S., Patterson, D., Schild, D., Featherstone, T. and Olson, M.V. (1991). Cloning and *in vivo* expression of the human GART gene using yeast artificial chromosomes. *The EMBO Journal* **10**: 1629-1634.

Green, E.D., Riethman, H.C., Dutchik, J.E., and Olson, M.V. (1991). Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11**: 658-669.

Green, E.D. Mohr, R.M. Idol, J.R., Jones, M., Buckingham, J.M., Deaven, L.R., Moyzis, R.K., and Olson, M.V. (1991). Systematic generation of sequence- tagged sites for physical mapping of human chromosomes: Application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics* **11**: 548-564.

Kwok, P.-Y, Gremaud, M.F., Nickerson, D.A., Hood, L., and Olson, M.V. (1992). Automatable screening of yeast artificial-chromosome libraries based on the oligonucleotide-liagation assay. *Genomics* **13**, 935-941.

Riles, L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., Braden, V.V., Depke, J.E., and Olson, M.V. (1993). Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6-kilobase pairs. *Genetics* **134**: 81-150.

Olson, M.V. and Green, P. (1993). Criterion for the completeness of large-scale physical maps of DNA. *Cold Spring Harb. Symp. Quant. Biol.* Vol. **58**: 349-355.

Gnirke, A., Iadonato, S.P., Kwok, P.-Y., and Olson, M.V. (1994). Physical calibration of yeast-artificial-chromosome-contig maps by RecA-assisted restriction endonuclease (RARE) cleavage. *Genomics* **24**, 199-210.

Gillett, W., Hanks, L., Wong, G. K.-S., Yu, J., Lim, R., and Olson, M.V. (1996). Assembly of high-resolution maps based on multiple complete digests of a redundant set of overlapping clones. *Genomics* **33**, 389-408.

Wong, G.K.-S., Yu, J., Thayer, E.C., and Olson, M.V. (1997). Multiple-Complete-Digest (MCD) Restriction-Fragment Mapping: Generating Sequence-Ready Maps for Large-Scale DNA Sequencing. *Proc. Natl. Acad. Sci. USA* **94**, 5225-5230.

Publications (Review articles, Book chapters, Essays)

Smith, R.P. and Olson, M.V. (1973). Drug-induced methemoglobinemia. Seminars in Hematology **10:** 253-268.

Olson, M.V. and Crawford, J.M. (1975). Semi-micro ion exchange in the freshman laboratory. *J. Chem. Educ*. **52:** 546-549.

Olson, M.V., Page, G.S., Sentenac, A., Loughney, K., Kurjan, J., Benditt, J., and Hall, B.D. (1980). Yeast suppressor tRNA genes. Transfer RNA: Biological Aspects (Soll, D., Abelson, J.N., and Schimmel, P.R., Eds.), pp. 267-279, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Olson, M.V. (1981). Applications of molecular cloning to *Saccharomyces*. Genetic Engineering, Vol. **3**, (Setlow, J.K. and Hollaender, A., Eds.), pp. 57-88, Plenum Press, New York, N.Y.

Carle, G.F. and Olson, M.V. (1987). Orthogonal-field-alternation gel electrophoresis. Methods in Enzymology, Vol. **155** (Wu, R., Ed.), pp. 468-482, Academic Press, San Diego, CA.

Helms, C., Dutchik, J.E., and Olson, M.V. (1987). A lambda DNA protocol based on purification of phage on DEAE-cellulose. Methods in Enzymology, Vol. **153** (Wu, R., and Grossman, L., Eds.), pp. 69-82, Academic Press, San Diego, CA.

Olson, M.V. (1989). Separation of large DNA molecules by pulsed-field gel electrophoresis. *J. Chromat*. **470:** 377-383.

Olson, M.V. (1989). Pulsed field gel electrophoresis. Genetic Engineering, Vol. **11** (Setlow, J.K., Ed.), pp. 183-227, Plenum Press, New York, N.Y.

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* **245:** 1434-1435.

Burke, D.T. and Olson, M.V. (1991). Preparation of clone libraries in yeast artificial-chromosome vectors. Methods in Enzymology, Vol. **194** (Guthrie, C., and Fink, G.R., Eds.), pp. 251-270, Academic Press, New York, N.Y.

Olson, M.V. (1991). The Human Genome Project and analytical chemistry: a tale of two cities. *Analyt. Chem.* Vol. **63**: 416A-420A.

Olson, M.V. (1991). Genome structure and organization in *Saccharomyces cerevisiae*. The Molecular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics (Broach, J.R., Pringle, J.R., and Jones, E.W., Eds.), pp. 1-39, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Olson, M.V. (1993). The Human Genome Project. *Proc. Natl. Acad. Sci. USA* **90:** 4338-4344.

Olson, M.V. (1995). A time to sequence. *Science* **270**: 394-396.

Funding received by Dr. Maynard V. Olson

Agency and Grant Number: NIH/NHGRI 5R01 HG01475
(01-02)
Grant Title: UW Genome Center-NIH
Dates of Entire Project Period: 04/11/96 - 06/30/98
Total Costs for Project: $4,305,197

Agency and Grant Number: NIH/NHGRI 5 R01 HG 01475
(03)
Grant Title: UW Genome Center-NIH98
Dates of Entire Project Period: 07/01/98 - 06/30/99
Total Costs for Project: $7,931,704

Agency and Grant Number: Cystic Fibrosis Foundation
Grant Title: CF - P. aeruginosa
Dates of Entire Project Period: 12/31/96 - 12/31/98
Direct Costs for Project: $1,000,000

Agency and Grant Number: Pathogenesis
Grant Title: Pathogenesis - P. a.
Dates of Entire Project: 12/31/96 - 12/31/98
Direct Costs for Project: $1,000,000

Pending

Agency and Grant Number: NIH
Grant Title: Discovering and Scoring
SNPs
Dates of Entire Project Period: 10/01/98 - 9/30/01
Direct Costs for Project: $1,997,712

Chairman CALVERT. Thank you, Doctor.

REASONS FOR FEDERAL GOVERNMENT TO COMPLETE HUMAN GENOME SEQUENCING

Chairman CALVERT. This question is first for Dr. Patrinos and Dr. Collins. Doctors, in a guest column in *The New York Times,* Dr. William Haseltine, a former Harvard Medical School professor and CEO of his own genomics company said the following, "It makes little sense for the Federal Government to go to the trouble of decoding the junk DNA. The $3 billion of federal money now devoted to the entire human genome should be spent instead on university-based research initiated by individual medical investigators. The era of government-sponsored big science in which a few laboratories receive as much as $10 million a year to analyze mostly junk DNA, while scientists doing disease-related research beg for financing should end."

At this point, if there is no objection, I would ask unanimous consent to insert the entire column at this point in this record and, hearing no objection, so ordered.

[The information referred to follows:]

Chairman CALVERT. And with that, I assume that each of you disagree and could you tell us why?

Dr. COLLINS. Every new development in science or in public policy tends to bring out of the woodwork individuals with fringe opinions who seek to take advantage of that new development to promote their own agenda. In this instance, the comments you quote are those of an individual who has a transparent financial conflict of interest in making such assertions, given that the future of his particular business enterprise would be best served by genome projects of all sorts, public or private, ceasing to exist. In addition, there are statements in those remarks which I think the vast majority, I would say greater than 99 percent of the scientific community, would profoundly disagree with. What Dr. Haseltine refers to as junk DNA includes sequences that play profound roles in juvenile onset diabetes, in cancer, in osteoporosis, and many other diseases and that has been scientifically demonstrated.

So, I would ask you not to consider that particular point of view as representative of the mainstream of scientific thought, either public or private.

Chairman CALVERT. Thank you for your clear answer. Dr. Patrinos. [Laughter.]

Dr. PATRINOS. I certainly couldn't have said it better myself.

REFOCUSING OF FEDERAL HUMAN GENOME PROJECT

Chairman CALVERT. Dr. Galas, in his testimony, and this, again, is for Dr. Patrinos and Dr. Collins, says the Federal Government approach should continue but should refocus its goals to produce a first draft, and that was indicated also by other witnesses, of the human genome as soon as possible. Will the government program consider this approach in collaboration with the private effort by Dr. Venter? Would you like to respond to that as well, Doctor? But, go ahead.

Mr. PATRINOS. As I mentioned in my oral remarks, this is, in fact, our intention. I agree wholeheartedly with what Dr. Galas has said about the value of providing this intermediate product as soon as possible and we certainly plan to deliver that intermediate product in coordination and in full cooperation with private-sector initiatives such as the initiative that Dr. Venter described.

Dr. COLLINS. It is, actually, worthy of note that there is a planning process under way right now for the NIH and the DOE genome programs. Ari and I work together on all of these planning processes and there was a meeting just three weeks ago involving more than a hundred scientists from various fields, most of them not genome scientists, to look at the next 5 years of the genome program. This subject of whether or not the publicly-funded effort should revise its strategy in light of the new developments was intensely discussed.

I think it's fair to say there is not complete unanimity on the answer to that question, in part, because of the uncertainty until that new initiative has moved forward a bit about exactly what it will look like. But I can certainly reassure you, this is being looked at with great intensity and I'm sure Ari would agree with me that as that data begins to become available we will be doing everything possible to adjust the strategy to make the most of that and to get to the goal as quickly as possible.

FEDERAL PROGRAM'S USE OF LATEST TECHNOLOGIES

Chairman CALVERT. Let's briefly discuss new technologies. There was discussion about that today also. Is the federal program using the latest technologies, for example, the new robotics advances in the last several years in our endeavor on our—answer the question, Doctor?

Mr. PATRINOS. There are indeed. Certainly among both our laboratory and academic performers in the human genome project there are many examples of cutting-edge technologies in robotics, sequencing technologies in general. This is a field, of course, that is rapidly changing. Advances are expected, as I mentioned earlier, and probably will be the norm rather than the exception, the surprising new developments, that is.

Dr. COLLINS. I would agree with that. In fact, I would add to it the federally-funded effort is not only using the new technology, we're developing a lot of it. The NIH component of the Human Genome Project spends $20 million a year on technology development. One of our successes is the DNA chip which was founded on the basis of a company that got going with an NIH grant about 4 or 5 years ago. So we are intensely interested in technology development. Many of our grantees are engineers, they are not necessarily all biologists, computer scientists, robotics experts, and the like. This is part of our goal.

Mr. PATRINOS. Let me add also one thing. At least the Department of Energy is investing some modest amount of funding in some of the cutting-edge technologies that we expect will be in place not in the next few years, but maybe 10–20 years from now, ones that are sort of blue sky right now with respect to their feasibility because we know that technology changes very, very quickly and how we sequence 20 years from now will probably be entirely different than how we are sequencing today.

Chairman CALVERT. Thank you. Mr. Roemer.

FEDERAL BUDGET FOR THE HUMAN GENOME PROJECT

Mr. ROEMER. Thank you, Mr. Chairman. I first of all want to thank the panel once again for your very helpful testimony on a very complicated subject. Certainly in my background in political science and in other areas that prepared me maybe better for running for Congress than it did contemplating many of these very complicated questions that you experts deal with, we're very appreciative for your, not only your expert testimony but, I think the way that you've also presented your testimony today as well too, in a very helpful, very persuasive, and very collaborative sense. We haven't had complete unanimity from the panel today and I want to get to that point. But first of all, Dr. Venter, I want to ask, to make sure that I heard your remark and clarify on it. You said that in this collaborative effort, you would not encourage Congress to cut the budget. In fact, you would encourage the Congress to increase the budget for this particular project, even though we're seeing this collaborative public/private partnership. Is that correct?

Mr. VENTER. That's absolutely correct, but not just for sequencing of humans. It is because we're going to have the sequence so much faster that we can now move to the phase that all of us hope to in the envisioning of the human genome project in the first place is starting to interpret and understand that genetic code. It will not be interpretable without having mouse and other genome sequences so the fact that human's going to be there faster, we need mouse even faster. Of the 60,000–80,000 human genes, there's only around 5,000 of those genes that have full-length cDNA sequences available to the worldwide community. Stepping up the effort so that every one of those human genes has a full-length cDNA sequence, which can be done on a very broadly-distributed effort in America's universities, we'll move forward to make sure we have the tools on a broadest possible sense for everybody to use. There's more reasons to fund more genomic research now than there ever has been.

Mr. ROEMER. So your testimony, which is very, you know, persuasive and compelling testimony, you say that in this collaborative effort, you are not replacing something that is being done in the publicly-funded research. In fact, in this collaborative effort, you are working together in a partnership and that does not mean that slices should be taken out of the existing budget.

Mr. VENTER. Well, we're certainly trying our best to work together and I don't think anything should be taken out of the budget. I've heard from some of my colleagues here that they've been criticized for wasting federal dollars based on this new announcement. I think that's a very unfair and unfortunate use of our announcement for people who have the agenda to attack the programs. I think it's a very different situation 3 years from now, perhaps looking back, if we are successful, and we would not have made this announcement if we didn't intend to be, but I think we, we want to be judged on our accomplishments, not by our press releases or announcements, and our accomplishments, hopefully, will show that it's wise to change the directions currently under way to work with us in a collaborative fashion to move this important research forward faster for everybody.

DR. OLSON'S CRITICISMS OF PRIVATE-SECTOR VENTURE

Mr. ROEMER. Thank you. Dr. Collins, you said in your testimony, I believe you said in your testimony, that you had worked at the University of Michigan and you had worked on the cystic fibrosis and Huntington disease structuring or the DNA researching and that that had taken close to a decade. You got some pretty strong criticism from Dr. Olson, even though you have some practical experience in academic life, he used pretty strong words such as this is science by press release, this is public policy by press release. He predicted there are going to be 100,000 gaps in the final product and misassembled data and so forth. How do you, as somebody that has been in his shoes in academic life at the University of Michigan, respond to this rather strong criticism and, well, let me leave it at that. And, I would just say that you certainly were not shy when it came to your remarks about Dr. William Haseltine's remarks as well too.

Dr. COLLINS. Mr. Roemer, I think there's a little confusion in the nature of Dr. Olson's remarks. Again, I'm the person who is responsible for overseeing the federally-funded effort at the National Institutes of Health on the genome project. I believe his comments about difficulties in assembling the structure were related to the announcement by Perkin-Elmer and Dr. Venter and not directed at the publicly-funded effort.

As a researcher who worked on cystic fibrosis and was fortunate to lead one of the two teams that worked together to find that gene, I can tell you that the 10 years that went by during that enterprise where I, as a physician, had to keep explaining to families whose children were increasingly getting sicker that we hadn't found the gene yet because it was just too hard, were among the more frustrating years of my life and I don't wish that on anybody in the future. And that is one of the major motivators to do this project and to do it right. Actually, Dr. Olson and I are pretty much in sync on this. I do believe that until the Perkin-Elmer effort has produced, over the course of the next 2 or 3 years, the data that will be required to evaluate this strategy, that exactly what kind of a product comes out of it is not knowable. It's not that we're just not doing our homework to know it, it's not knowable. It is a problem that hasn't been tried before and, therefore, I agree with Dr. Olson that the publicly-funded effort, which Dr. Patrinos and I are responsible for, should not drastically alter our strategy which is targeted toward having this final complete, highly-accurate product until we have some more data.

Mr. ROEMER. But I'm asking you objectively as a scientist to comment on Dr. Olson's remarks about Dr. Venter, that's what the question was about, not a confusion as to where the criticism was coming from—or where it was directed.

Dr. COLLINS. I think as, as I tried to say, that this approach to putting together the human genome sequence is bold. It is of uncertain success value. It could be that 2 or 3 years from now, as Dr. Olson is predicting, we end up with a rough draft which is actually rough enough that it is very difficult to work with. The publicly-funded effort is probably the only part of this enterprise that's absolutely dedicated to obtaining the completely contiguous, highly-accurate, close-all-the-gaps enterprise and I think we need to take that responsibility and take it seriously and will continue to do so. But I welcome this new initiative and look forward to seeing what's going to happen. It's a scientific experiment; we like that. Scientists are energized by the opportunity to see a new approach tried out. It will take a while to find out, but that's what science is all about.

Mr. ROEMER. So, you are consistent in your initial enthusiasm with your testimony for Dr. Venter's efforts; however, you do have concerns as a scientist as to what it may produce. You may not agree with some of Dr. Olson's conclusions, but you are saying that first of all, this effort should go forward; secondly, you are excited about the potential; thirdly, you do have questions as Dr. Olson does about what the outcome may be?

Dr. COLLINS. I think every scientist has to agree with Dr. Olson when he says show me the data; then I will make up my mind.

ETHICAL, LEGAL AND SOCIAL CONCERNS

Mr. ROEMER. Dr. Patrinos, you said in your initial testimony as well, that you're excited, you support this collaborative effort. You also said that you have some ethical and legal and social concerns. Can you be a little bit more specific as to what those might be and do they come back to some of Dr. Olson's concerns about access, privacy, or any of those other issues?

Mr. PATRINOS. Of course, as you know the Human Genome Project from the very beginning identified the ethical, legal, and social implications of this project as very important, in fact the HGP carved a significant piece of the budget from the very beginning to deal with those issues and that's something we've been doing, Dr. Collins and I, for quite some time. My comment was mostly made in the context of the faster delivery of the product. In a sense the faster delivery of the product will confront us with many of the ethical and legal and social implications of the project that have been articulated by many of the scientists and the science managers involved——

Mr. ROEMER. Please give me some examples of what that conflict——

Mr. PATRINOS. Issues of privacy and confidentiality of genetic information, issues of insurance and employment discrimination, the multitude of issues in forensics. You know the list is endless, we can have an entire hearing solely devoted to this as I'm sure Dr. Collins would be delighted to have such a hearing because this is one of his very important private concerns. So I was making reference to the issue of having that information faster than perhaps we had expected a few years back and, thus, forcing us to confront some of these issues sooner rather than later.

Mr. ROEMER. I would hope that our Chairman might be amenable to having another hearing on that and learning of some of those potential problems and gleaning maybe some of the potential answers to those problems and maybe having an ethicist as well to discuss what those might be. With that, I understand my colleague from Michigan has to leave the hearing and I'd be happy to yield back the time, although I'm sure the Chairman has been very patient with me and I don't have any time left, so.

Chairman CALVERT. Well, we certainly can come back for another round, so, that's not a problem. Mr. Ehlers.

Mr. EHLERS. Thank you, Mr. Chairman. It is a very interesting hearing. I apologize for being a little bit late, but it's been one of those days again. It all sounds terribly complicated to me, then maybe because I'm a physicist I am used to dealing with simple problems, just electrons and nucleae and quarks, and so forth.

PATENTABILITY OF HUMAN GENOME

Mr. EHLERS. Dr. Venter, I think I understand the difference between your approach and what we may call the standard approach but I'm interested in your comment that you, in your written testimony you say that this will, your actions will basically make the human genome unpatentable. Can you explain that to me? Are you saying that you are going to wipe out so much of it that, and you're not planning to patent it, that no one will be able to, or what? Just what do you mean by that?

Dr. VENTER. Well, our plan, as we've announced in our so-called press barrage was that we do plan to make the sequence data we generate over the next couple of years on the complete human genome accessible to the public. We do not plan to patent that human genome sequence, the human chromosomes, or the complete genome. In fact, by putting it in the public domain as the individuals who sequence that information, if we do not patent it, we will be making it and rendering it unpatentable by others. However, we will be using that sequence as the beginning for discoveries, as all others will be able to, once we release it to discover new genes that are key for pharmaceutical development, new hormones that could become pharmaceutics themselves and the key to understanding key human diseases.

Some of those genes, such as the gene for human insulin when Genentech patented it, that allowed the process to begin for human insulin to be available to diabetics as a drug because someone was willing to produce it. We will be patenting cDNA's in a limited number for new, exciting discoveries that we make with the genome. The human chromosome sequence itself and the human genome will be unpatented by us and because we will be doing this so quickly, we are going to render it unpatentable by others.

DIFFERENCE BETWEEN FEDERAL HUMAN GENOME PROJECT AND PRIVATE-SECTOR VENTURE

Mr. EHLERS. Let me ask another question. I've done some experiments which demand extreme precision, parts and 10 to the 9th, and very, very careful work over some time. I've also done some which are called quick and dirty where you are just trying to outline the parameters of something to decide whether or not there is something worth investigating there. Is that, in a sense, the difference between the so-called human genome project and your work?

Dr. VENTER. Absolutely not. In fact, I appreciate you asking that question. Quick does not mean dirty. Quick means better technology, better approaches, new strategies. We're going to be sequencing the human genome 10 times. The sequences that we've done in the past are some of the most accurate sequences ever put in the public domain by any scientist and we're going to have the same standard for the sequences that we do with the human genome. It's a completely different strategy; in fact, we think it's a scientifically more justifiable strategy than relying on clones that have been processed several times, coming from limited parts of the genome, not necessarily reflecting the entire genome. We're starting with the entire set of human DNA, the entire set of chromosomes and using that and going right into the sequencing machines to generate the data. We're relying on new algorithms we've developed, new strategies we've developed, and the very forefront of computing to be able to reassemble all these pieces into the genome.

Mr. EHLERS. So your statement would be that your method is going to yield results with the same completeness and the same accuracy as the Human Genome Project?

Mr. VENTER. We actually feel that our approach is going to yield more completeness and at least the same level of accuracy as done by the best groups, including our own that have now been sequencing the human genome by the existing strategy. It is unknown, you know, my colleagues are correct in characterizing this as an experiment. But some of these same individuals are the same ones that criticized our approach to sequence the hemophilus influenza genome. In fact, one of the questions I get asked most often is why didn't we just apply to the Federal Government for funds to do this new strategy.

Well, I think it's clear, Maynard Olson is the Chairman of that review committee and I think you've heard the comments. I think if we went and asked for $300 million to do this new project, that they might get some good chuckles out of it, but it's not the way new initiatives can be made.

Mr. EHLERS. So, basically, what I hear you saying is it's not the contrast between the precise, complete experiment and the quick-and-dirty experiment but rather the contrast between a bureaucratic risk-free approach and a more thoughtful modern approach.

Mr. VENTER. I think that would characterize my view quite well.

RECAPTURING PRIVATE INVESTMENT

Mr. EHLERS. All right. Next question. You mentioned $300 million. If you are putting $300 million in, obviously, you hope to get a return on that, or at least your investors do. How will you recapture your investment?

Mr. VENTER. Well, the goal, in fact, the strategy that we're taking proves our philosophy that getting the sequence is only the first step. And while we feel morally compelled to release that genome sequence to the entire public, and the companies that have proceeded on the basis of secrecy are taking things very much in the wrong direction, the business strategy is going to be building the ultimate genome database relating every bit that we can of the human genome information out to individuals, to physicians, to biotech companies and pharmaceutical companies. On the other side, and one of the things that comes out of this whole genome strategy that hasn't been discussed, is we get the sequence from both chromosomes, both alleles, and we're, in the first three months of operation, going to have over 3 million polymorphic variations that we're going to use as the basis for setting up high throughput screening of patients, of individuals, in part for the pharmaceutical industry as a basis for the new clinical trials stratifying patients. This is going to be the basis of the future of individualized medicine and we feel we can build a very major business without relying on secrecy and allowing other people to use the same sequence, discoveries, for their businesses and for their own scientific discoveries.

Mr. EHLERS. Thank you. I find this very interesting and, as Dr. Collins observed, this is an experiment. I will be very interested in seeing the results of the experiment and it will be fun to get you back in about 3 or 4 years and read your prepared testimony and your answers back to you at that point.

Mr. VENTER. Thank you. I appreciate that.

Mr. EHLERS. And find out who really was out on this one. Thank you very much.

Mr. VENTER. Thank you.

Chairman CALVERT. Mr. Ehlers. Mr. Bartlett.

Mr. BARTLETT. Thank you very much, and I apologize for not being able to be here for the testimony.

TENSION BETWEEN FREE MARKET AND WIDE INFORMATION DISSEMINATION

Mr. BARTLETT. We obviously, as a society, have two objectives that are in tension here. One is the objective to make knowledge of the genome widely available so it will benefit the maximum number of people. The other is to use competition which, wherever it's used in our free market society makes the product or the service better and it makes it cheaper. And, obviously these two things tend to be in tension here. How do we proceed so that we maximize the contributions that competition will make and, yet, be assured that we are going to have as wide a possible dissemination of this information so that there will be the maximum benefit from it?

Mr. VENTER. I assume that question is for me?

Mr. BARTLETT. Well, for whoever.

Mr. VENTER. Okay. Well, we're going to be disseminating our information, first in terms of the raw sequence itself will be provided to the world for free and also the world will have access to this new database that we're building. We're not here to try to persuade NIH or DOE or anybody else not to do what they are doing. We're not concerned with competition. I would hate to see the federal budget cut because of the basis of what we're doing. I think we can proceed much better if we work together. There's clear complementary approaches taken with both strategies that will yield a much more complete, faster product, even sooner than we could possibly anticipate. We would like to be judged, as I said earlier, on what we accomplish. We're not concerned with competition, other than my concern is as a scientist who first spent 10 years at the NIH and before that 10 years trying to get NIH grants, my institution is totally funded by NIH, DOE, NSF, and Department of Defense grants. I have as much concern for the public funding of science as I do for the private funding of science and if it goes in the wrong direction, we all lose from that proposition.

Dr. COLLINS. Could I add a comment? I think you asked a very appropriate question about how to balance these two forces, but I think this is a very good example where those two forces actually are synergistic on both counts. Having a public/private partnership of this sort should speed up getting the final product, that's the nature of a synergism, a collaboration, if it works, and we are determined to see that it does work. But I believe having the public effort continue to be vigorously involved in this as much or more so than they have been, is also the best insurance that the data is made publicly accessible. I do not question for a moment Dr. Venter's sincerity in his statement that this data will be made available on a quarterly basis in a database that anybody can look at. I know that that is what he is committed to doing. But, after all, the sequence of the human genome is of such profound importance, that I think a scenario where large quantities of it were only available within the database of a single private entity might be a rather unstable situation. If business demands were to change or personnel were to change or the stockholders were to decide it's not such a good thing to be giving this all away anymore, one would not want to see a circumstance where the publicly-funded effort was suddenly found to have dropped the ball. We don't intend to drop the ball.

Mr. BARTLETT. Thank you. I am very supportive of private-sector funds in this kind of scientific endeavor. Our federally-funded scientific organizations have done an exemplary job through the year, through the years, but in spite of that, I have a growing concern that when you have put all of your eggs in this basket which is controlled by a Congress which can, which can change course very quickly, that we put the future of science at risk. And so I am very supportive of any mechanism which attracts more private-sector funds and more competition. I think that whenever you have all of the direction of a program under the control of a single entity, in this case, ultimately the Congress, I think that you, that you buy some risk that you don't need to buy, if the ventures are broadly supported through competitive infusion of private-sector funds. So, thank you very much for your answers.

Chairman CALVERT. Thank you, Mr. Bartlett. When you say things change rapidly, everything except this Congress. Mr. Roemer, do you have any concluding questions?

CONCERNS ABOUT PUBLIC ACCESS TO INFORMATION

Mr. ROEMER. Yes, Mr. Chairman, just one or two, and I appreciate getting into a second round here. I'm reading from a *Washington Post* article, Tuesday, May 12, 1998, and in it, I quote Dr. Olson saying, "Even though there are promising public access," and I guess you mean Dr. Venter's group?

Mr. OLSON. I haven't read that article——

Mr. ROEMER. "They control the terms and there is a history of terms being more onerous than is acceptable to most scientists." Is that your quote?

Mr. OLSON. I haven't seen the article in question, but——

Mr. ROEMER. Does that sound like your quote?

Mr. OLSON. Sounds like me. [Laughter.]

Mr. ROEMER. Can you clarify what you meant by that quote and maybe we can get Dr. Venter to respond to that?

Mr. OLSON. Well, as I say, it would help if I had a little more context, but, at the close of my written——

Mr. ROEMER. Let me try to help you there, Dr. Olson, because I'm not sure if, you know, in a newspaper article, they're limited by space and I'm not sure how they can provide in terms of the lead-in. The previous paragraph says, "These companies have been granted scores of patents on their genetic discoveries raising fears among some critics that a handful of companies will control the commercialization of a vast and potentially lucrative biological resource. Those fears arose again yesterday when Venter announced his new project," then your quote.

Mr. OLSON. I see, well, at the close of my written testimony, I actually encouraged the Congress to keep careful track of the impact of intellectual property issues, particularly on basic research which is my interest. And I do encourage you to do so. I share Congressman Bartlett's view that this dynamic involvement of multiple sectors is critical to the health of contemporary science.

My own interest happens to be in, my most vital interest happens to be in the public sector, and I think what I was referring to there, in the short history of proprietary databases, and these databases, which are privately funded are at their inception proprietary and should be proprietary, they're paid for by private funds, that there is a history of the data being made available to academic investigators only in return for what are sometimes called reach-through agreements in which subsequent discoveries made by academic investigators using those data will be, the intellectual property status of these subsequent discoveries will be influenced by the agreement that must be signed at the time that the data are made available. And I think I was simply trying to make the point in this context that there are different degrees of accessibility and I think most scientists are comfortable, particularly with genome sequence data, that it be absolutely unimpeded by hidden costs.

Mr. ROEMER. So your reference of onerous, terms more onerous than is acceptable to most scientists, would refer to these reach-back provisions——

Mr. OLSON. Yes.

Mr. ROEMER. That are sometimes used. Dr. Venter, I want to give you time to respond to that. You say in the next paragraph that with the exception of perhaps 100 to 300 genetic sequences that you expect will show special commercial promise, the company will make all the genetic information available free to the world's scientists. You say, I quote, excuse me, you said, and I quote, it would be morally wrong to hold the data hostage and keep it secret, unquote. Is it morally wrong to keep the 100 to 300 genetic sequences from this same kind of scrutiny or providing this to the scientific community?

Mr. VENTER. Well, as Dr. Olson knows from his own work on the *pseudomonas originosa* genome with private companies, there is a big difference between secrecy and accessibility. One hundred percent of the sequence that we will generate will be publicly available. We will be putting it in the public domain. Having intellectual property rights on specific genes have no impact on Dr. Olson or anybody else. They allow whatever company has those rights the ability to commercially produce that product, whether it be insulin or raythocroeatin, whether key drugs that have a tremendous impact on human health.

I agree with Dr. Olson's concerns about reach-through rights and we've made that a key tenet of our philosophy. In fact, putting the human sequence in the public domain guarantees that there are no rights, reach-through or otherwise, that come with this. Any licensing that we do will not have reach-through rights. We're basing this company and the commercial aspects on this on building the best database ever. If it's not, nobody will pay to have access to it because they won't want it. If we can't measure polymorphisms faster and better and more meaningfully than anybody else, we won't make money. If the genes we discover don't have an impact on medicine, nobody will want to license those. None of those have any impact whatsoever on whether the fundamental data is widely and freely available to others.

CONSEQUENCES OF INTELLECTUAL PROPERTY/PATIENT/PRIVACY RIGHTS

Mr. ROEMER. Finally, Dr. Collins, let me just end with this final question and I'm not sure that I will phrase it the way that I want so bear with me. Is there, then, a difference here that we're speaking about in this collaborative effort that if Dr. Venter's group sequences the DNA, does the DNA sequencing for some form of cancer, or Parkinson's, or Alzheimer's and has a patent or privacy on that, is there different access, then, for that particular scientific knowledge than there would be under the research that the NIH and DOE are doing? And what are the consequences of that?

Mr. COLLINS. These are subtle and difficult questions, but let me do the best I can. The way that the publicly-funded effort is going forward is that we insist that our grantees, who are working at universities all over the country and also at the DOE labs (and this also applies on the international scene to the large-scale genome sequencing efforts that are going on in other countries) deposit their sequence data within 24 hours of the time it reaches an assembly of 2,000 letters in a row or more.

We are not, at the NIH, allowed to deny our grantees the opportunity to file for intellectual property rights on things they discover with NIH funds, because of the Bayh-Dole Act. So, we cannot tell them not to do that, but by insisting upon this early deposit of the data, the net outcome of that seems to be that that filing is not going on.

To our knowledge, none of the genome centers are filing for intellectual property protection. They just don't have time and their goal is, really, to get the data out there so that other scientists can figure out what's there. So, they are pouring out data every day of this sort for the rest of the scientific community to use, to analyze, to try to figure out. Is there a cancer gene in yesterday's output from the St. Louis center? Is there a diabetes gene in the day-before-yesterday's output from Maynard Olson's Center at the University of Washington? It takes another set of steps to figure that out.

The sequence itself is publicly accessible. It is truly in the public domain. "Public domain" is usually reserved to say there has been no intellectual property placed upon this, so the sequence is both publicly accessible and it is in the public domain. Now future investigators, who figure out the value of a particular gene sequence, may learn that it causes a particular disease or learn that it can be turned into a pharmaceutical, and then may decide that they have added enough value to that to meet the criteria of novelty, nonobviousness, and utility and file a patent on it. Those investigators might be in academia or they might be in companies, and the Patent and Trademark Office then decides whether they've made a convincing case or not.

Mr. ROEMER. Thank you. I think each time you ask a question, it begs some more questions. It's been a fascinating panel and you've been very helpful and I hope we can do another panel like this and add to some more questions. And I appreciate the Chairman, your foresight in having this hearing today.

Chairman CALVERT. Thank you, Mr. Roemer.

CONSEQUENCES OF PRIVATE-SECTOR VENTURE FOR FEDERAL HUMAN GENOME PROJECT

Chairman CALVERT. I have just a quick question for Dr. Olson. Obviously you are a skeptic when it comes to the private sector initiative described here today. If this project is likely to fail, in your estimation, should we just ignore it and continue the federal program that we have today unchanged?

Mr. OLSON. Well, I want to make clear that failure is a relative term. I have emphasized that I believe it will produce a huge amount of extremely useful data. I don't believe that it will meet the quality standards which have been outlined. And I think that the federal program would be well advised over the next 2 or 3 years to concentrate on defining the cost-benefit tradeoffs associated with the high-quality sequence product. No known approach is going to produce a perfect product. Indeed, perfect is not well-defined in the context of intrinsically variable structure like the human genome, but I believe that the federal, the unique niche for the federal program over the next few years is to refine the methods that are required to produce the best available product that can be achieved at a reasonable cost, and I would define a reasonable cost as roughly current levels of funding.

One of the difficulties in this highly-collaborative model, which is certainly correct in principle, but a technical point about the proposed Perkin-Elmer strategy is that it is heavily back loaded in terms of answering my concerns. Even a simple theoretical analysis of this approach to sequencing the genome, indicates that particularly this issue of gaps, will only be addressable relatively late in the project. One simply can't tell from the early indicators how that issue is going to go.

So, I believe the federal project should focus on the high quality and the definition of high quality, the exploration of the cost/benefit issues, the demonstration that by fail-safe methods we can produce such data over the next few years and when this rather back loaded information comes to us from this initiative or other initiatives, all I can really say is that we will look at it very closely and I'm certainly pleased to hear these renewed strong assurances that we'll be able to look at it. That is the data that will be there.

Chairman CALVERT. Thank you, Doctor.

EFFICIENCY OF FEDERAL HUMAN GENOME PROGRAM

Chairman CALVERT. Dr. Galas, you've got some experience in government, now in the private sector. How would you evaluate the efficiency of the government program and their ability to make changes as technology improves?

Mr. GALAS. I, actually I think that the human genome program, perhaps because of the fact that, unlike most federally-supported programs there's internal competition of a friendly type within the program having two agencies running it actually has been very responsive in being able to take advantage of new technologies. With the DOE and the NIH looking over each other's shoulders, I think actually the human genome program has done reasonably well in that regard. I'm sure it could be improved and I'm sure they are constantly looking at how to do so, but I think they can take advantage of that.

I would say that, if I might address some of the comments that Dr. Olson just made, I think that in fact there probably does exist a strategy that would be a different strategy from what is being right now in the program. Maybe only slightly different, but different nonetheless, that does not, on the one hand, depend entirely on the success or the back loaded success of the private-sector program but can take advantage of data as it's released from this program and enhance the federal effort, but not depend on the success of the private program, but merely be accelerated by it if it does succeed. And I think that's what the federal program should focus on, rather than focusing on the downstream, final product which I think, quite frankly, that Dr. Olson makes when he talks about sequence quality on the one hand and scientific standards on the other, they are not equivalent at all. Those are really not, that's an inequality that can't be made I think.

I think there's a rational strategy in there which does have a continually improving quality of sequence, or a staged quality of sequence that would get some of the fundamental, really important biological data out sooner and benefit us, be able to take advantage of what data is released by the private sector without making any assumptions about either the quality or whether or not they'll succeed.

Chairman CALVERT. Thank you.

Mrs. LEE. No questions? Any other questions from the panel?

I want to thank our witnesses for very interesting testimony and answers to our questions. I think you can rest assured, I doubt very much if Congress will cut funding on the Human Genome Project and we look forward to a successful conclusion and certainly, Doctor, we wish you well in your new venture. Thank you.

[Whereupon, at 2:40 p.m., the hearing was adjourned.]

[The following material was received for the record.]

APPENDIX 1: Answers to Post-Hearing Questions Submitted by Members of the Subcommittee on Energy and Environment

COMMITTEE ON SCIENCE
SUBCOMMITTEE ON ENERGY AND ENVIRONMENT
U.S. HOUSE OF REPRESENTATIVES

Hearing
on

The Human Genome Project:
How Private Sector Developments Affect the Government Program

June 17, 1998

Post-Hearing Questions Submitted to

Dr. Aristides A. Patrinos
Associate Director of Energy Research for
Biological and Environmental Research
U.S. Department of Energy
Washington, DC

Post-Hearing Questions Submitted by Chairman Calvert

Scientific Justification for Completing Government-Funded Sequencing of Entire Human Genome

Q1. Critics of the government program say that sequencing the entire human genome is a waste of the taxpayer's money. Please explain why it is scientifically necessary to complete the entire process.

A1. We estimate that the human genome, approximately 3 billion bases of DNA, contains about 80,000 genes. It has been estimated that the DNA sequence (cDNAs) containing the specific instructions for making these 80,000 protein products may occupy only about 3% of the total genome. While the specific role for the remaining 97% of the genomic sequence is unknown at this time there is no way at present to reliably recognize in advance those components that we need to sequence. Even if we could physically recognize the important sequences there is no method to select out in an economical way, those parts that are biologically significant for sequencing. Merely sequencing the expressed cDNAs certainly won't deliver the needed information to understand human biology—on this there is very strong agreement from the research community. For example, essentially all of the information that is critical for the proper regulation of genes, information vital to the proper "turning on" and "turning off" of genes so that they become operational at the right times and in the right cells is not recovered in the expressed cDNAs. Damage in these regulatory regions has been shown to be an important cause of genetic disease in humans.

We can and must do the best job we can to prioritize what we sequence so that, in our estimation, we are getting the best value for the money. However, we need to know the entire sequence to fully explore the complexity of human biology and fully exploit the information in the human genome.

Efficiencies of DOE's Joint Genome Initiative vs. Three Different DOE Laboratory Programs

Q2. In your testimony, you describe the Joint Genome Initiative, which allows for joint management and oversight of three different laboratory programs, those at Lawrence Berkeley, Lawrence Livermore and Los Alamos. The JGI was implemented seven years into the program. Were there inefficiencies and higher costs as a result of separate management of the three labs' programs and, in hindsight, would it have been better if joint management existed from the beginning of the program?

A2. The first phase of the Human Genome Program (HGP), closely coordinated between the DOE and the NIH was the phase of exploration requiring many independent pursuits. Also, it was necessarily devoted to laying the groundwork for the intensive sequencing effort that has begun in the last couple of years. In 1990, at the start of the HGP, sequencing technologies were not advanced enough, nor efficient enough, to accomplish the task of sequencing 3 billion base pairs at the expected funding levels and in the expected time frame. Additionally, large scale chromosomal mapping efforts were undertaken to provide the detailed physical maps that it was thought would be critically necessary to achieving the complete genome sequence. Each of the three DOE Lab genome centers carried out parallel and non-overlapping research efforts to map different chromosomes and to explore technologies that would accelerate the sequencing. Not until the genome project was ready to switch directions to full scale production sequencing, was the nature of the task such that issues of critical mass, economies of scale, and sharpness of focus together made central management the correct paradigm.

Post-Hearing Questions Submitted by Democratic Members

Difference Between the DOE-NIH and "Shotgun" Human DNA Sequencing Approaches

Q1. How does the DOE-NIH approach, projected to be completed by the year 2005, differ from the Venter-Perkin-Elmer plan to use the "shotgun" method to sequence the human genome in three years?

A1. The DOE/NIH commitment is to produce a complete and accurate image of the human genome by 2005. In the first 2 years (FY 1997 and FY 1998) of the production effort, the approach taken insisted on full sequencing accuracy, high continuity, and detailed mapping (location) knowledge every step of the way, in part to ensure that these meritorious standards could be achieved at affordable cost. This assurance now being in hand, DOE is considering an approach that we produce an intermediate draft version of the genome based on a "mapped clone shotgun method"—in contrast to the "whole genome shotgun method" being followed by Venter-Perkin-Elmer. In the mapped clone shotgun, in which we shotgun sequence, but only within already mapped clones that are about 1/20,000 the size of the genome, we can have a much higher assurance of positional and sequence assembly validity than the Venter-Perkin-Elmer method. In practice, the two approaches will complement each other and be extremely useful to the scientific community.

Role of DOE and NIH in Collaboration with Private-Sector Venture

Q2. Do you see a role for DOE and NIH to collaborate with Venter and Perkin-Elmer to complete sequencing of the human genome?

A2. Yes, a very significant opportunity exists. In practical and scientific terms, the two approaches can strongly and synergistically complement each other. In fact, the clone resources that Venter-Perkin-Elmer will utilize have been developed and made available to the public by DOE and NIH; and the DOE is funding projects that will provide the sequence information from the ends of 600,000 BACs (bacterial artificial chromosomes) that will form the scaffold needed for linking the human genome sequence together in the Venter-Perkin-Elmer Plan. The DOE and NIH will help both private and the public sequencing efforts by aggressively completing the BAC-end sequence set, as well as developing a high resolution radiation hybrid map of BAC ends and other sequence markers, and the mapping of all cDNA ESTs (Expressed Sequence Tags) against the BAC libraries being sequenced.

Q2.1. How would this be done?

A2.1. On the Venter-Perkin-Elmer side, prompt and complete sharing of their raw data with the public is the core requisite of making the two efforts mutually complementary. On the public side, it is necessary that DOE and NIH simultaneously produce a high quality, fully mapped, draft ('scaffold') intermediate version of the genome, on top of which the Venter-Perkin-Elmer sequence could most usefully be assembled (adding depth for improved accuracy and coverage). The public effort would then proceed to complete this jointly constructed draft version to full coverage and accuracy sooner than originally planned and at a lower cost.

Q2.2. At what stage would it be done?

A2.2. The Venter-Perkin-Elmer venture has projected a completion date of two to three years; thus, to be effective, any collaborative elements need to be in place quickly and ongoing during the course of the project. As mentioned above, some of the needed efforts are already underway and it is anticipated that the remaining components will be initiated before January 1999.

Concerns of International Collaborators About Intellectual Property Rights and Patenting

Q3. The international Human Genome Organization (HUGO) has been fairly vocal about their feelings concerning intellectual property rights and patenting.

Q3.1. How have the international collaborators responded to this proposed venture?

A3.1. With very serious concern. These concerns derive from the immense and essentially unrestrained possibilities that exist for intellectual property rights control when extremely high rate, highly automated data generation techniques are used by a privately owned company to produce and combine both "composition of matter" information (sequence data) with "utility" information (e.g., mapping and gene expression data), to form the basis of patent applications *en masse*. Thus, the response to this venture by the Wellcome Trust in Great Britain, the principal public funder of human genome sequencing efforts at the Sanger Center in Britain, was to announce that they would double the budget in support of human genome sequencing at the Sanger Center. The Sanger Center, like its US counterparts has a policy of daily release of sequence.

Q3.2. How do you plan to allay their concerns that the race for patenting will (1) hinder information exchange and (2) result in unnecessary and costly duplication?

A3.2. (1) The DOE and NIH must not deviate from their clearly stated policy, elaborated at a series of meetings of the heads of sequencing programs and large sequencing labs in the US and other countries, of nightly electronic release of newly determined human sequence, without any restrictions on availability.

The Venter-Perkin-Elmer group has publicly stated that the vast majority of sequence information that they determine will be deposited in public databases within a few months of sequencing. The several hundred genes that they say they will focus on represents much less that one percent of all human genes. Thus information exchange for the vast majority of human genes should not, theoretically be compromised by this private sequencing effort. Similarly, there should not be a costly race for patenting for >99% of the human genes simply as a result of this one private effort. It should not be surprising, however, that "use patents" for human genes may become a significant issue when large numbers of human genes are finally identified, whether by private or public methods.

(2) As mentioned earlier, the Venter-Perkin-Elmer genome sequencing efforts are seen by DOE as complementary and not duplicative of the public efforts by the US public Human Genome Program. With regard to the public efforts, the Human Genome Organization (HUGO) is coordinating, through a Web site, a current view of which centers/labs are sequencing which human chromosomes or chromosome fragments. This site is accessible to anyone via the Web. The purpose of this HUGO effort is to minimize duplication among publicly funded sequencing efforts.

COMMITTEE ON SCIENCE
SUBCOMMITTEE ON ENERGY AND ENVIRONMENT
U.S. HOUSE OF REPRESENTATIVES

Hearing
on

The Human Genome Project:
How Private Sector Developments Affect the Government Program

June 17, 1998

Post-Hearing Questions Submitted to

Dr. Francis S. Collins
Director, National Human Genome Research Institute
National Institutes of Health
U.S. Department of Health and Human Services
Bethesda, MD

Post-Hearing Questions Submitted from Chairman Ken Calvert

Scientific Justification for Completing Government-Funded Sequencing of Entire Human Genome

Q1. Critics of the government program say that sequencing the entire human genome is a waste of the taxpayer's money. Please explain why it is scientifically necessary to complete the entire process.

A1. The more we study DNA, the more we understand how it carries out its amazing work. Genes affect almost all important biological processes, at least in part. This includes those processes that lead to or are involved in disease. By identifying the gene(s) associated with a disease, we will gain important understanding that can help us develop therapies or preventive strategies. The Human Genome Project, including sequencing the entire human genome, is designed to speed up the process of gene identification and make it much more cost-efficient. Genes, we have learned, are made up of several parts that control their activity. Sometimes all the parts are clustered in the same DNA neighborhood, but other times, the parts may be scattered far apart from each other. Also, at times mistakes in DNA spelling in regions thought to be of no importance turn out to contribute to disease risk. We already have found such examples for cancer, diabetes, and osteoporosis. Some important parts are very easy to spot and some aren't. Knowing all of the parts of a gene is critical to understanding how it works. Many of the other approaches to gene identification that have been used so far cannot find all of the parts of every gene (that is one reason why these other approaches tend to be somewhat faster and appear to be less expensive). Having a complete genome sequence is the only way to find all of the parts of all of the genes that may affect human health. The Human Genome Project will provide a truly complete genome sequence containing no gaps. That level of completeness we believe is necessary to provide researchers with the best possible tool for understanding the function of genes and their role in human health and disease.

Post-Hearing Questions Submitted by Democratic Members

Difference Between the DOE-NIH and "Shotgun" Human DNA Sequencing Approaches

Q1. How does the DOE-NIH approach, projected to be completed by the year 2005, differ from the Venter-Perkin Elmer plan to use the "shotgun" method to sequence the human genome in three years?

A1. Sequencing was once done by hand as a series of chemical reactions—a slow and costly method. Now, machines can read the sequence quickly, but current instruments can only read short DNA fragments at a time. So, using a strategy referred to as "shotgun" sequencing, an investigator randomly cuts DNA into small fragments. These fragments are small enough for sequencing machines to read. Then, the scientist must correctly reassemble all of these sequenced fragments in order to properly reconstruct the full-length DNA sequence. The reassembly of this giant puzzle is carried out largely by highly skilled scientists using sophisticated computer programs.

The sequencing strategy the public genome project uses employs shotgun sequencing of DNA fragments that have been carefully mapped and catalogued. This strategy is designed to maximize the accuracy of reassembling the sequenced fragments, because the scientist knows where the fragments belong. Even so, the scientists periodically encounter DNA regions that are particularly difficult to sequence, and which therefore require special attention. Because all the fragments have been catalogued, a scientist can return to these difficult spots after most of the genome has been sequenced and assembled to work on closing the gaps and strengthening the weak areas so that the entire sequence will, in the end, be finished to very high quality. The international sequencing community, whose goal is to complete the human DNA sequence by 2005, has agreed to a policy of releasing completed sequence every 24 hours into a free, publicly-accessible database. More than 10 percent of the human sequence is now available in a public database, and about half of that is already "finished."

The sequencing strategy proposed by scientists at Perkin-Elmer, Inc. and Dr. Venter also employs shotgun sequencing, but differs from the public effort in several significant ways. First, that strategy, called "whole-genome shotgun sequencing", employs fragments that have not been previously mapped or catalogued. Because the scientist does not know where in the morass of 3 billion base pairs the fragment might belong, the task of reassembling the fragments becomes far more difficult. Many believe, this difficulty in reassembly will inevitably lead to many gaps and misassembled regions in the sequence. These scientists believe that, on its own, the quality of the "whole genome shotgun sequence" will not be as high as that planned for the publicly-funded sequence. For example, when a scientist encounters a fragment that is particularly difficult to sequence, he or she will not be able to return to the fragment later because it has not been catalogued. The Perkin-Elmer-Venter approach does not propose to fill in all the gaps left by these unsequenced fragments, thereby creating a product that may be incomplete for many research uses. Not having a sequence of the highest quality will be a serious problem when the gaps and errors occur in DNA regions with biological significance.

In addition, release of sequence data from the Perkin-Elmer-Venter effort will occur quarterly, rather than daily. Although the company states that sequence will be made public, release will be significantly slower than data release from the publicly-funded effort. As a result, the larger research community's access to this valuable data will be slowed down. Furthermore, the new company maintains the right to patent the most biologically important gene data.

Role of DOE and NIH in Collaboration with Private-Sector Venture

Q2. Do you see a role for DOE and NIH to collaborate with Venter and Perkin-Elmer to complete sequencing of the human genome?

Q2.1 How would this be done?
Q2.2 At what stage would it be done?

A2. Partnership with the private sector is both necessary and desirable and we welcome this new initiative by Perkin-Elmer and Dr. Venter. In the year ahead, we will look carefully at the ways in which this private initiative and the publicly-funded effort can be complementary. If need be, the federal effort is fully prepared to adjust its strategy. In fact, in late May, just weeks after the private sector announcement, there was a meeting involving more than 100 scientists from various fields and from both the public and private sectors, to look at the next five years of the genome project. The subject of how collaboration might occur and whether or not the publicly-funded effort should revise its strategy was intensely discussed. I think it is fair to say there is not yet complete unanimity on the answer to those questions. The Perkin-Elmer/Venter proposal is a scientific experiment; we like that. Scientists are energized by the opportunity to see a new approach tried out. It will take time, at least 12 to 18 months, to develop enough data to allow the usefulness of the approach to be evaluated, and to assess the quality of the product, but that is what science is all about.

Concerns of International Collaborators About Intellectual Property Rights and Patenting

Q3. The international Human Genome Organization (HUGO) has been fairly vocal about their feelings concerning intellectual property rights and patenting.

Q3.1 How have the international collaborators responded to this proposed venture?
Q3.2 How do you plan to allay their concerns that the race for patenting will (1) hinder information exchange and (2) result in unnecessary and costly duplication?

A3. On May 13, 1998, the Wellcome Trust announced their intent to increase its support of British science in the sequencing of the human genome. Previously, the Wellcome Trust had committed to funding the sequencing of one sixth of the human genome at the Sanger Centre in the United Kingdom. The May 13 announcement, doubled that commitment to one third of the genome and expressed concern with regard to a number of aspects of the private sector initiative. In the press release accompanying the announcement, the Wellcome Trust stated:

> “The Wellcome Trust has today announced a major increase in its flagship investment in British science in the sequencing of the human genome.... The Trust is concerned that commercial entities might file opportunistic patents on DNA sequence. The Trust is conducting an urgent review of the credibility and scope of patents based solely on DNA sequence.... This week a commercial venture announced its intention to produce partial sequence of the human genome, to delay release of this information and to have exclusive rights to patent some of these sequences.... The Wellcome Trust believes that the human genome should be sequenced, through an international collaboration, as speedily and accurately as possible, with the results being placed immediately in the public domain.”

The Wellcome Trust is the leading European funder of human genome sequencing. Its early support of work in the field has enabled Dr. John Sulston, Director of the Sanger Centre, and his colleagues, to generate one third of all the human sequence which had been produced at the time of the May 13 announcement.

With regard to patenting, this is a difficult area that does not lend itself to simple answers. The way the publicly-funded effort in the United States, which includes HGP grantees from universities all over the country and also at the DOE labs, is going forward is that we have agreed with our international sequencing collaborators to deposit sequence data within 24 hours of the time it reaches at least an assembly of 2,000 bases, or letters, in a row. Absent a finding of exceptional circumstances, we are not at the NIH allowed to deny our grantees the opportunity to file for intellectual property rights on things they discover with NIH funds, because of the Bayh-Dole Act. As a practical matter, however, the publicly supported sequencing community has agreed to a 24 hour data release policy, and we are not aware that there have been any patent filings.

Therefore, the sequence itself is publicly accessible. It is truly in the public domain, which usually is reserved to say there have been no intellectual property restrictions placed upon the data. So, future investigators, who figure out the function of a particular gene sequence and/or turn that sequence information into a pharmaceutical or a new diagnostic, may decide they have added enough value to meet the patent criteria of novelty, nonobviousness, and utility, and file for a patent. Those investigators may be in academia, here in the United States or abroad, or they might be in private industry. But all seeking patent protection must make a case sufficient to convince the Patent and Trademark Office that their discovery deserves protection under the law.

Federal Government's Cost to Completely Sequence the Human Genome

Q4. Dr. Collins, you have indicated that to date, the Federal Government has spent about $100 million on human genome sequencing. How much more do you think it will cost the Federal Government to completely sequence the human genome using the federal sequencing approach?

A4. The original projection was that the entire Human Genome Project, including mapping, sequencing, technology development, model organisms, informatics, and ELSI would cost $200 million a year for 15 years, for a total of $3 billion in 1990 equivalent dollars. If you include the FY'99 budget request, a total of $1.5 billion in 1990 dollars will have been spent over a 9 year period. This is approximately $300 million below the $1.8 billion originally projected for the Project over the first 9 years. So we are significantly under the projected cost of the Project.

Up to this point, the Project has only spent about $100 million on human production sequencing. Now it is a very critical question, what will it cost the government to completely sequence the human genome? The difference between 50 cents per finished base and 49 cents per finished base is $30 million worth of cost. Greater reductions in the per finished base cost will yield more significant reductions in cost.

The NHGRI has instituted a new method of bringing together our genome sequencing centers. They have agreed to cooperate to share their technology ideas and to figure out who is saving money and at what step or steps in the process. The NHGRI also will continue to support research to improve sequencing technology and reduce costs.

I think it is a little hard to predict how things will go in the next 6 or 7 years, particularly with regard to the impact on costs of further developments in technology and activity in the private sector. But I am very optimistic that the sequencing component of the Project can be accomplished within the projected budget. To date, we have met our goals on time, and under budget. I would hope the Human Genome Project in the future will be judged by the total budget that was required to provide a highly accurate, publicly accessible, contiguous, finished sequence as soon as possible.

COMMITTEE ON SCIENCE
SUBCOMMITTEE ON ENERGY AND ENVIRONMENT
U.S. HOUSE OF REPRESENTATIVES

Hearing
on

The Human Genome Project:
How Private Sector Developments Affect the Government Program

June 17, 1998

Post-Hearing Questions Submitted to

Dr. J. Craig Venter
President and Director
The Institute for Genomic Research
Rockville, MD

Post-Hearing Questions Submitted by Republican Members

Will the Private Initiative Duplicate the Federal Human Genome Project?

Q1. Please tell us, should your initiative be successful, will you in fact have duplicated the federal program, or, as some have said, given us a "synopsis" of the human genome?

A1. By obtaining the complete DNA sequence of the human genome by the year 2000, our new venture will make the science of genomics directly applicable to combating human disease in the broadest way possible. We won't duplicate the federal program because we'll actually obtain the complete sequence and make it available before that effort is complete. We will, however, be building our program on resources and strategies that have been developed as a result of the federally-funded initiative. As I indicated in my testimony, obtaining the complete sequence of the human genome is not an end to itself, but represents a beginning for the real research that will allow us to better understand the disorders that afflict humankind. The federally-funded program needs to be positioned to ensure this new research takes place, whether in the year 2005 as previously planned or in the year 2000.

Concern About Release of Data to the Public

Q2. In his testimony, Dr. Francis Collins expressed concern that your plans to release data to the public on a quarterly basis is not sufficient. Please tell us your response to that.

A2. As a requirement for receiving a grant from either the Department of Energy or the National Human Genome Research Institute for DNA sequencing the recipients are required to release sequence data as soon after it is generated as possible. This is a requirement for publicly-funded activities. As I indicated in my testimony, we don't presume to be able to understand the biological significance of all the data that we will generate in completing the sequence of the human genome. As scientists, we also understand the importance of sharing data. The current model that is employed by most commercial organizations in this field is to keep human DNA sequence data private. We intend to share the data that we generate on a quarterly basis. There are obviously people and organizations, especially in the public sector, who don't feel this frequency is adequate. However, we are not required to meet the objectives of the publicly-funded project and given the current commercial alternative we believe our approach is very appropriate.

Recommendations for Restructuring the Federal Human Genome Project

Q3. In your testimony, you say the impact your venture will have on the federal program will be to re-orient it to focus on research into the genetic impact of disease on a broad basis. Could you please elaborate on that and tell us any specific recommendations you have on how the federal program should be restructured.

A3. The Human Genome Project is about much more than just obtaining the complete human DNA sequence. The sequencing is just the biggest initial hurdle that needs to be cleared. Once the human sequence is complete, the information will exist to begin in-depth research into the actual functioning of the genetic code. One critical resource that will be required to undertake this task will be providing researchers access to full-length cDNA clones. This will allow researchers to study specific genes in great detail and at this time there is no resource for this material. Only a small percentage of the genome is actually made up of genes, but these regions will attract a significant amount of the initial research activity from both private and public entities. However, there will be real value in understanding all aspects of the human genome, and NHGRI is a logical place to undertake this activity.

Post-Hearing Questions Submitted by Democratic Members

Availability of Genomic Information to the Scientific Community

Q1. Although details of your business venture with Perkin-Elmer Cooperation may not be finalized, you and Tony White, Chair, President, and Chief Executive Officer of Perkin-Elmer, have indicated that you intend to make genomic information from this venture available to the scientific community. How can we be assured that this will happen?

A1. On June 20, 1997, The Institute for Genomic Research (TIGR) and Human Genome Science (HGS) ended a collaborative arrangement that required TIGR to forego payments totalling $38 million. The primary reason for my choosing to end this relationship and access to significant financial resources was a philosophical disagreement about the public release of DNA sequence data. The day after this relationship was terminated, TIGR made the largest deposit of DNA sequence data into the public domain in history. When I entered negotiations with the Perkin-Elmer Corporation to undertake this new venture, the first point of agreement was the requirement that human genome data would be made publicly available. If agreement had not been reached on this point, we would not be discussing this new venture. I don't know of many organizations that would forego $38 million to ensure that DNA sequence data would be made publicly available, and this act should provide a high-level of comfort to you and others that this data will be made available to the public.

Timeliness of Release of and Compensation for Human DNA Sequence Data

Q2. Once obtained, how soon and for what economic compensation will this information be released by your new company?

A2. As previously indicated, the human DNA sequence data will be made publicly-available at no charge on a quarterly basis for the scientific community. The details and pricing models for the new venture's products are still being determined at this time.

Plans to Patent Genomic Sequences

Q3. Obviously, you and the Perkin-Elmer Corporation plan to patent a number of genomic sequences.

Q3.1 Since the patenting criteria include utility, in addition to novelty and unobviousness to peers, will the sequences you plan to patent correspond to particular biological functions or genetic traits?

Q3.2 Your past patenting attempts involved these expressed sequence tags (ESTs) you discussed in your testimony. To the best of my knowledge, these requests were denied. Could you explain to me (1) why that was and (2) what in your current EST strategy will allow for the patenting of these tags.

A3. As you correctly noted, the NIH chose to file patents for the ESTs identified by my lab. This initial application was rejected and NIH chose not to appeal the ruling. We are not planning to seek patents on broad sets of ESTs similar to what was done at NIH. Instead, we plan to fully characterize a small subset of key genes for which we will seek to identify and understand their biological significance. In an article published in the May 1, 1998 issue of *Science,* John Doll, Director of Biotechnology Examination at the U.S. Patent and Trademark Office (PTO), indicated that the same patentability analysis which is conducted for any other application will be conducted in the area of genomics. It is our intent to satisfy the PTO standards for those discoveries on which we seek to file for patents. I have attached a copy of that article for your information.

Uniqueness of Expressed Sequence Tags

Q4. How unique are these tags in terms of their ability to identify an expressed gene or locate a gene on a larger map of the genome. Is it a 1:1 correspondence in terms of ONE tag corresponding to a ONE part of the genome? What does that tell us about the functional purpose of that gene?

A4. There is generally a 1:1 to correspondence between an EST and its location on the genome. With regard to functionality, it depends upon what else we know about the EST as to whether it indicates any specific function. For example, if a human EST matches a sequence from another organism and there is some function associated with it, then it is likely the sequence will have a similar function in humans.

Role of DOE and NIH in Collaboration with Private-Sector Venture

Q5. What do you see as DOE and NIH's role in collaboration with yourself and Perkin-Elmer?

Q5.1 How would this collaboration be done?
Q5.2 At what stage would it be done?

A5. NIH, DOE and the new venture could establish the basis for collaboration nearly immediately, and to some degree we already have. As I indicated in my testimony, certain resources that have been publicly-funded like bacterial artificial chromosomes (BACs), will provide the framework for assembling the genome data that we will generate. As we publicly release DNA sequence data, this data will be available for all DOE and NIH grantees to use in their research.

There are more specific areas of collaboration that could be undertaken that have been discussed on a preliminary basis. One area of particular significance that I have spoken about with Dr. Varmus is that of the ethical, legal, and social implications of the genomic research. A number of concerns have been raised in the past few years about issues relating to genetic testing, discrimination in insurance, and privacy of individual genetic information. These issues and other issues will only become more important in the coming years, especially as we speed up completion of the sequence of the human genome. NIH has set aside a portion of its annual funding to address these issues, and this is an important and logical area for collaboration. I intend to follow-up on my conversation with Dr. Varmus to identify specific activities which we can jointly undertake.

Restrictions on Researchers' Ability to Obtain Human DNA Sequence Information

Q6. What restrictions will be placed on researchers' ability to obtain this information?

A6. The human DNA sequence information will be made publicly available to researchers on a quarterly basis. There will be no restrictions placed on this data by the new venture.

Relation of New Venture to the Federally-Funded Human Genome Sequencing Effort

Q7. How will you and Perkin-Elmer executives relate your program to the federally funded human genome sequencing effort? To the efforts of other biotechnology companies?

A7. The new venture that we are undertaking, if successful, will advance the efforts of all human genome research activities. All programs either publicly or privately funded will gain some advantage by utilizing the information encoded in the entire human genome. We hope to work with all researchers to improve understanding into the genetic basis of disease and to one day assist in the creation of therapeutics that will improve human health.

COMMITTEE ON SCIENCE
SUBCOMMITTEE ON ENERGY AND ENVIRONMENT
U.S. HOUSE OF REPRESENTATIVES

Hearing
on

The Human Genome Project:
How Private Sector Developments Affect the Government Program

June 17, 1998

Post-Hearing Questions Submitted to

Dr. David J. Galas
President and Chief Scientific Officer
Chiroscience R & D Inc.
Bothell, WA

Post-Hearing Questions Submitted by Chairman Calvert

Practical Value of Federal Completion of Entire Human Genome Sequencing Process

Q1. In your testimony, you say that, even if the Federal program agrees to the "first draft" approach you recommend, it should then go on to complete the entire sequencing process. Please tell us the practical value this will have.

A1. The "first draft" approach will make available valuable information that can be used to locate genes and certain other important tasks for projects currently being pursued in the public and private sectors. It is important, as I testified, that this information be available as soon as possible to help advance a wide range of present and planned research work – thus the value of the "first draft". Researchers will use this information to provide clues to enable them to do further work, including more detailed sequencing, in specific places in the genome of direct interest. In no way, however, should this "first draft" be viewed as the final result of the genome project. The complete sequence information is needed in any case to provide a complete picture of the biological function of the genome. When the final product is available in the databases any further sequencing by researchers will not be necessary, and even more time and resources will be saved than with their use of the "first draft" data.

Post-Hearing Questions Submitted by Democratic Members

Impact on Current Efforts

Q1. How would your current efforts be affected by the joint venture?

A1. If the joint venture succeeds as planned, we would welcome the new data that will be available in the databases, and use it as soon as it is available. Our efforts will thus be enhanced by the joint venture.

Importance of Genomic Data That May Be Withheld

Q2. How important do you feel the 100 to 300 sequences that would be withheld are to the broad assemblage of knowledge?

A2. Since many companies now withhold the results of their own proprietary work on genes, including their identity and function, I doubt if this will change the landscape to a significant degree. I am confident that any withheld genes will be discovered in short order in the course of normal efforts by the federal program or by other academic or industry researchers. I would expect that any gene withheld in this way would result only in a short delay in its availability to the rest of the community.

Reasonable Fees and Conditions to Private-Controlled Genetic Information

Q3. Could you share with the committee what you feel are reasonable fees and conditions to the genetic information Perkin-Elmer will control.

A3. Unfortunately it is too early for me to make reasonable estimates of this. It depends on the specific information (which is highly variable in its value to the commercial sector) and the context of the state of knowledge at the time when it would actually be made available.

Rights of Individuals—Privacy and Compensation Issues

Q4. Please discuss rights of individuals whose specific genomic sequences could lead to a commercially successful drug? Are there privacy issues? Are their fair compensation issues?

A4. Use of individual's DNA should only be done under fully informed consent, which should include the use of genetic information for research purposes. While there are strong privacy issues that, in my view, must be dealt with clearly and carefully, in my opinion, individuals should have no rights to research information that is gained by using a biological sample as part of a research program. Any future claims to completely unknowable future results that their sample may be used to produce should be explicitly renounced ahead of time in the informed consent process by the individual. The advance of medical science helps all of us and our future descendents. This is part of the fair compensation for cooperation in a research study of any kind, including one that involves genetics.

COMMITTEE ON SCIENCE
SUBCOMMITTEE ON ENERGY AND ENVIRONMENT
U.S. HOUSE OF REPRESENTATIVES

Hearing
on

The Human Genome Project:
How Private Sector Developments Affect the Government Program

June 17, 1998

Post-Hearing Questions Submitted to

Dr. Maynard V. Olson
Professor of Medical Genetics and Genetics
Department of Molecular Biotechnology
and
Director, Genome Center
University of Washington
Seattle, WA

Post-Hearing Questions Submitted by Democratic Members

Concerns About Ability to Access Genomic Information

Q1. Do you have concerns about your ability to obtain access to genomic information that may come out of this new venture? If so, what are they? Are you aware of any past or current problems in this area?

A1. I have concerns in two areas. First, current promises about data release cannot be regarded as binding commitments. The public position taken by Perkin Elmer is that there will be excellent access to all the data. However, the business interests of the firm will be constantly re-evaluated in the years ahead. Perkin Elmer is free, as it should be, to change its position. Secondly, much of the utility of the data to experts will depend on access not just to processed data, but also to the raw output from the instruments. The amount of raw data will be vast and it will require pro-active effort on Perkin Elmer's part to insure that these data are accessible in a readily analyzed form. Since it is difficult to see why Perkin Elmer will have any incentive to make the needed effort, accessibility is likely to become bogged down in haggling with federal agencies about who will pay for and take responsibility for the data handling and whether or not the cost is justified.

Impact on Current Efforts

Q2. How would your current efforts be affected by the joint venture?

A2. The answer to this question depends on how it goes. Right now the only effect is that it has generated inordinate amounts of discussion for which there is not much basis. If the effort actually results in quick delivery of a high-quality human sequence, it would have a major effect on my activities: I could move on a few years earlier than planned to other research goals. However, I will only contemplate such a move once I see that the venture is really fulfilling the strong claims that have been made for it. My expectation is that the venture will end up having only a minor effect on my activities. Scientists are always making minor adjustments to rapidly changing external developments. It will have more impact on scientists who are in the thick of analyzing particular problems in human genetics (as opposed to engaging in large-scale genome analysis). These scientists will benefit from earlier access to valuable data than they would otherwise have been the case.

Importance of Genomic Data That May Be Withheld

Q3. How important do you feel the 100 to 300 sequences that would be withheld are to the broad assemblage of knowledge?

A3. As long as all the data are released, as promised, and there is no effort to deter academic researchers from using these data in follow-up studies, I am unconcerned about whether Perkin Elmer attempts to patent 100 genes or 100,000 genes. It is not up to scientists to write or interpret the patent law. I only become concerned when intellectual-property issues become an obstacle to the free pursuit of new knowledge.

Reasonable Fees and Conditions to Private-Controlled Genetic Information

Q4. Could you share with the committee what you feel are reasonable fees and conditions to the genetic information Perkin-Elmer will control.

A4. I assume that this question concerns licensing fees to commercial firms who want to use information that is protected through patents or copyrights. I have no expertise in this area. My opinion, expressed as that of a scientist rather than an expert in the commercial aspects of biotechnology, is that it does not serve the public interest for pharmaceutical companies to confront a tangle of expensive licensing issues whenever they choose to pursue a new product-development program. Most of the real costs and real difficulies associated with drug development lie far downstream from DNA sequencing, and the rewards of successful drug-development efforts should be kept well aligned with the steps in the process that involve the highest risk and require the largest investment.

Rights of Individuals—Privacy and Compensation Issues

Q5. Please discuss rights of individuals whose specific genomic sequences could lead to a commercially successful drug? Are there privacy issues? Are their fair compensation issues?

A5. This area bears watching. Certainly, there are privacy issues whenever DNA sequences go into databases. I believe that all such data should meet a high standard of anonymity, and we should also avoid drifting toward, just as a matter of convenience, obtaining a high proportion of human sequence from the DNA of a small number of individuals. In general, the tradition of obtaining research samples from individuals who are largely motivated by altruism--with compensation that is only related to the time and effort that they must expend in providing the samples--serves the public interest well.

Biomedical research depends on ready availability of enormous numbers of research samples acquired from patients and volunteers, under conditions of informed consent, every day. It would not serve the public interest to inject legal contracts and commercial agreements into the relationship between research subjects and researchers. We also do not want to turn the process into a lottery. Any particular commercially important discovery can be traced to a particular sample or small number of samples; however, in most cases, the individuals who provided those samples are no more deserving of special rewards than the thousands of other people who also allowed their samples to be used for similar research purposes.

In short, we should insist on high standards of privacy, anonymity, and informed consent but should not start a system in which donors of research samples have an ongoing legal and commercial interest in the research projects that employ their samples. However, sticky issues will still arise, particularly when the special commercial potential of a particular sample can be recognized in advance of extensive scientific analysis or when samples are collected in cultural settings where the research subjects have had little exposure to modern medicine or do not feel they benefit from advances in medical knowledge. Nonetheless, the more closely we can stick to a system in which well informed research subjects volunteer to provide research samples out of altruistic motives, the better the public interest will be served.

APPENDIX 2: Additional Materials for the Record

POLICY: GENOMICS

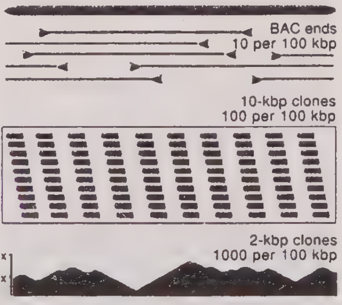
Shotgun Sequencing of the Human Genome

J. Craig Venter, Mark D. Adams, Granger G. Sutton,
Anthony R. Kerlavage, Hamilton O. Smith, Michael Hunkapiller

The Human Genome Project (HGP) was officially launched in the United States on 1 October 1990 as a 15-year program to map and sequence the complete set of human chromosomes and those of several model organisms. The HGP is laying the groundwork for a revolution in medicine and biology. Its importance is underscored by the level of funding from the National Institutes of Health, the Department of Energy (DOE), the Wellcome Trust, and other governments and foundations around the world.

From the inception of the HGP, major technical innovations that would affect its timetable and cost were considered essential to success. The development of bacterial artificial chromosomes (BACs) (*1*) provided a key advance. BACs are propagated in *Escherichia coli* and carry large [~150–kilobase pairs (kbp)] inserts stably. In contrast, ordered cosmid clones that served as the basis of yeast (*2*) and *Caenorhabditis elegans* (*3*) genome sequencing projects are less stable and much shorter (~35 kbp). Fluorescent labeling of DNA fragments generated by the Sanger dideoxy chain termination method has been the mainstay of almost all large-scale sequencing projects since the introduction of the first semi-automated sequencer by Applied Biosystems in 1987 and the development of Taq cycle sequencing in 1990. New models of the sequencer that can process more samples, Taq polymerase engineered especially for sequencing, and higher sensitivity dyes have improved throughput, accuracy, and operating costs. Publication of the first genome from a self-replicating organism, *Haemophilus influenzae*, was based on a whole-genome shotgun (random sequencing) method (*4*). A set of algorithms called the TIGR Assembler (*5*) together with scaffolding sequences from both ends of 18-kbp inserts in bacteriophage lambda clones were critical for determination of correct order and assembly. Eight additional genomes have since been completed by these methods (*4*, *6*, *7*), and several others are nearing completion, including genomes with high GC (~65%) and high AT (~82%) composition, which present special problems for sequencing and assembly.

Current approaches to human genomic sequencing rely on building sequence-ready maps over regions ranging in size from hundreds of kilobase pairs to whole chromosomes and then sequencing individual BACs spanning these regions through a combination of shotgun and directed approaches. This method can produce highly accurate sequence with few gaps, although most sequencing centers have encountered regions that appear to be unsequenceable by current technology. The up-front steps of building and validating the sequence-ready map and subclone library construction and the downstream steps of directed gap filling are generally considered to be rate limiting. About 120 Mbp of human genomic sequence were completed through 1997, and another 200 Mbp are planned for 1998.



Covering the genome. A 100-kbp portion of the genome showing expected clone coverage.

The recent announcement by Perkin-Elmer of a new, fully automated sequencer (ABI PRISM 3700) permits a reevaluation of strategies for completing the human genome sequence. This instrument is a capillary-based sequencer that can process ~1000 samples per day with minimal hands-on operator time (~15 min compared with ~8 hours for the same number of samples on ABI PRISM 377s). This reduction in operating labor, coupled with automation of sample purification and sequencing chemistry enabled by the sequencer's improved detection sensitivity, suggests that the tens of millions of sequencing reactions necessary to complete the human genome can be performed more quickly and at lower cost than previously anticipated. The Institute for Genomic Research (TIGR) and Perkin-Elmer have started a program to complete this task within 3 years using this new technology and a whole-genome shotgun strategy that obviates the need for a sequence-ready map before sequencing. We intend to form a new company to carry out this venture and develop a commercial business based on these efforts. The cost of the project is estimated to be between $200 million and $250 million, including the complete computational and laboratory infrastructure to develop the finished sequence and informatics tools to support access to it.

The whole-genome shotgun strategy involves randomly breaking DNA into segments of various sizes and cloning these fragments into vectors. The presence of repeat elements, regions that are unclonable in a particular vector, and the benefit of having more DNA available in clones than is actually sequenced (see figure and table) require that multiple vector libraries be used. A library of pUC18-based plasmids containing ~2-kbp inserts will provide most of the sequencing templates. These clones will be sequenced from both ends to produce pairs of linked sequences representing ~500 bp at the ends of each insert. End sequences from a library of low–copy number plasmid clones containing ~10-kbp inserts will provide medium-range linking, including spanning the common Line-1 and THE repeat elements. Use of multiple cloning systems should help to reduce the effect of sequences that are unclonable or otherwise not present in one of the libraries. The goal is to generate 70 million high-quality DNA sequences totaling ~35 billion bp (10× coverage) of raw human sequence.

An argument for whole-genome shotgun sequencing of the human genome was made (*8*) and rebutted (*9*) in 1997. A year later, we see developments in technology and a new resource for this project consisting of a large database of end sequences of BAC clones. This will provide a framework for linking contigs over larger regions. Currently, the DOE is funding a program at TIGR and the University of Washington to sequence both ends (~500 bp from each end) of 300,000 human BAC clones. This BAC-end sequencing strategy was originally proposed to accelerate genome sequencing by providing markers every 5 kbp throughout the genome (*10*).

The new human genome sequencing facility will be located on the TIGR campus

---

J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, and H. O. Smith are at The Institute for Genomic Research (TIGR), Rockville, MD 20850, USA. M. Hunkapiller is at Perkin-Elmer Applied Biosystems, Foster City, CA 94404-1128, USA.

in Rockville, Maryland, and will consist of 230 ABI PRISM 3700 DNA sequencers with a combined daily capacity of ~100 Mbp of raw sequence. The facility will also have the infrastructure to produce ~100,000 template preps and ~200,000 sequencing reactions daily. This includes both custom and off-the-shelf robotic devices for picking colonies, pipetting, and thermal cycling. Quality control and assessment procedures will be implemented at each stage of the process.

Accompanying the challenge of obtaining the primary sequence data in a rapid and cost-effective way is the major challenge of assembling raw data into contiguous blocks (contigs) and assigning those to the correct location in the genome. Complete contiguity of the clone map should theoretically be achieved by about 9× coverage, so the 46× coverage (see table) allows for substantial deviation from the statistical model. The pairs of end sequences from each template are constrained by the assembly algorithms to be directed toward one another in the final assembly and located at a given distance apart depending on the insert size of the originating library. Although the BAC end sequences will be the primary scaffold onto which the end sequences from the smaller clones will be assembled, other available resources will be used to verify the alignments and place contigs on individual chromosomes. The most important of these resources is the large number of sequence tagged site (STS) markers that constitute the physical maps that have been produced by many laboratories during the first phase of the HGP. There currently are about 45,000 STS sequences, including about 30,000 that are well ordered along the chromosomes and provide a defined marker approximately every 100 kbp (*11*). Expressed sequence tags (ESTs) that tag 50 to 80% of human genes (*12*) and full-length cDNA sequences spanning up to 5 Mbp of genomic sequence will be used to verify the final assemblies. There are likely to be contigs that are misassembled or incorrectly linked together because of the presence of long, duplicated segments of the genome. We expect to recognize and correct ambiguous or conflicting assembly structures using a combination of manual inspection and directed experimental effort.

The aim of this project is to produce highly accurate, ordered sequence that spans more than 99.9% of the human genome (*13*). The 10× sequence coverage means that the accuracy of the sequence will be comparable to the standard now prevalent in the genome sequencing community of fewer than one error in 10,000 bp. It is likely that several thousand gaps will remain, although we cannot predict with confidence how many unclonable or unsequenceable regions may be encountered. We look forward to working with other genome centers to ensure that the sequence meets the requirements of the scientific community for accuracy and completeness; this will include making clones and electropherograms available.

An essential feature of the business plan is that it relies on complete public availability of the sequence data. The four primary business areas are high-throughput contract sequencing, gene discovery, database services, and high-throughput polymorphism screening. A major consequence of the analysis of data generated by this project will be the creation of a comprehensive human genomic database. It will contain an extensive set of DNA and protein features derived from the primary sequence. DNA features will include identified genes and their regulators, repeats, links with genetic and physical mapping data, synteny with other species, and polymorphisms. Because of the importance of this information to the entire biomedical research community, key elements of this database, including primary sequence data, will be made available without use restrictions. In this regard, we will work closely with national DNA repositories such as National Center for Biotechnology Information. We plan to release contig data into the public domain at least every 3 months and the complete human genome sequence at the end of the project. We also envision providing at a minimum connect fee online access to these data and many of the informatics tools to interpret them. We will also market the database system to commercial companies engaged in pharmaceutical and biotechnology research.

Vector type	Insert size (kbp)	Number of		Coverage (×)	
		Clones	Sequences	Sequences	Clones
High-copy plasmid	2	30,000,000	60,000,000	8.5	17
Low-copy plasmid	10	5,000,000	10,000,000	1.4	14
BAC	150	300,000	600,000	0.1	15
Total		**35,300,000**	**70,600,000**	**10**	**46**

Analysis of coverage. As each clone is not completely sequenced, there is a greater coverage of clones than sequences in the assembly. We assume a 500-bp average read length and 3.5-Gbp genome size.

Because the whole-genome shotgun approach will contain data from multiple individuals (the exact number has not yet been determined), we will generate a large number of precisely located single-nucleotide polymorphic (SNP) sites spanning the genome. Using technology being developed at Perkin-Elmer, we will generate assay systems to validate these markers and select a highly informative set of at least 100,000 SNPs. We plan to work with commercial partners to screen DNA samples associated with diseases or other conditions in an effort to link them with particular genetic loci. The assay systems will also be marketed by Perkin-Elmer to third parties for in-house research. Although we do not plan to seek patent protection for the randomly selected SNPs, we may seek patents on diagnostic tests based on the association of particular SNPs with important phenotypic traits.

We also do not plan to seek patents on primary human genome sequences. However, we expect that we and others will be able to use these primary data as a starting point for additional biological studies that could identify and define new pharmaceutical and diagnostic targets. Once we have fully characterized important structures (including, for example, defining biological function), we expect to seek patent protection as appropriate. Given both the complexity and scope of the information contained in human genome sequence, as well as its public availability, we would expect to focus our own biological research efforts on 100 to 300 novel gene systems from among the thousands of potential targets. If we are successful in these efforts, the patents would be available for licensing to interested parties.

Although it is clear that shotgun sequencing at this scale has never been attempted, it is our hypothesis that the desired result is achievable. While building the human genome sequencing infrastructure we plan to attempt to demonstrate the effectiveness of the shotgun strategy on a large and complex genome, in collaboration with Gerald Rubin (Howard Hughes Medical Institute/University of California Berkeley) and the Berkeley *Drosophila* Genome Project (BDGP). *Drosophila melanogaster* represents a good system for testing the whole-genome shotgun strategy because of the extensive physical and genetic maps that exist, the presence of about 12% of the genome as high-quality finished sequence with which to compare shotgun assembly results, and its importance as a model organism. We will work fully with the BDGP to facilitate the final closure process (which includes making clones and electropherograms available), with the expected result being a highly accurate and contiguous set of chromosome sequences. The *Drosophila* genome sequence will be deposited in GenBank both while in progress and at completion. An international workshop is being organized for September 1998 to develop a plan for completing the *Drosophila* genome that encourages participation of all groups currently working on this project.

It is our hope that this program is complementary to the broader scientific efforts to define and understand the information contained in our genome. It owes much to the efforts of the pioneers both in academia and government who conceived and initiated the HGP with the goal of providing this information as rapidly as possible to the international scientific community. The knowledge gained will be key to deciphering the genetic contribution to important human conditions and justifies expanded government investment in further understanding of the genome. We look forward to a mutually rewarding partnership between public and private institutions, which each have an important role in using the marvels of molecular biology for the benefit of all.

References and Notes

1. H. Shizuya *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89,** 8794 (1992).
2. A. Goffeau *et al.*, *Nature* **387** (suppl.), 5 (1997).
3. J. Sulston *et al.*, *ibid.* **356**, 37(1992).
4. R. Fleischmann *et al.*, *Science* **269**, 496 (1995).
5. G. G. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Technol.* **1**, 9 (1995).
6. C. M. Fraser *et al.*, *Science* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); J.-F. Tomb *et al.*, *Nature* **388**, 539 (1997); H.-P. Klenk *et al.*, *ibid.* **390**, 364 (1997); C. M. Fraser *et al.*, *ibid.*, p. 580; C. M. Fraser *et al.*, *Science*, in press.
7. D. R. Smith *et al.*, *J. Bacteriol.* **179**, 7135 (1997); G. Deckert *et al.*, *Nature* **392**, 353 (1998).
8. J. Weber and E. W. Myers, *Genome Res.* **7**, 401 (1997).
9. E. Green, *ibid.*, p. 410.
10. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
11. T. Hudson *et al.*, *Science* **270,** 1945 (1995); C. Dib *et al.*, *Nature* **380,** 152 (1996); G. D. Schuler *et al.*, *Science* **274,** 540 (1996).
12. M. Adams *et al.*, *Science* **252**, 1651 (1991); M. Adams *et al.*, *Nature* **37** (suppl.), 3 (1995); L. Hillier *et al.*, *Genome Res.* **6**, 807 (1996).
13. Considerable sequence will be generated from regions of heterochromatin including centromeres, telomeres, and ribosomal DNA arrays, which are not targeted by HGP sequencing laboratories. We will make unique assemblies where possible in these regions.

BOOKS: PALEOBIOLOGY

Palaeobiography

Paul Copper

Life. A Natural History of the First Four Billion Years of Life on Earth. RICHARD FORTEY. Knopf, New York, 1998. xiv, 347 pp., + plates. $30 or C$42. ISBN 0-375-40119-9.

A portentous book title as bold as this—*Life*—is bound to raise a few eyebrows. It is also almost certain to catch the eye of the book browser. In a drama bolder and more sweeping than *Gone with the Wind*, Richard Fortey sketches the full story of life on Earth, the stage and the actors, over more than four billion years. Originally published in Britain as *Life: An Unauthorized Biography* (Harper Collins, 1997), this bright brown volume, plastered with the imprint of *Archaeopteryx* (the oldest known bird), is as encompassing as its title suggests. Fortey, senior palaeontologist at the Natural History Museum, London, takes us on a roller coaster from the spawning of the simplest unicellular organisms during violent infancy of the Earth; through monumental crustal upheavals, voyages of continents, and mass extinctions; to an ending at the dawn of human-recorded history.

The key to this book, a layperson's guide to the secrets of fossils and environments most ancient, is the way the author has magically transposed and integrated his academic biography and intellectual growth into the natural history of life. I know of no other "autobiography"—if the book can be called one—quite like this, where the author's life is stitched into such an immense stretch of time. Neatly and adroitly, Fortey weaves his personal observations, his encounters with scientists (famous and less well known), and his introductions to controversies (century-old and contemporary) into a chronological tapestry of life on Earth.

The text literally begins with *Salterella*, the vessel that in 1967 carried Fortey, then a young Cambridge undergraduate, to his first field season in Spitsbergen. *Salterella* is also one of the oldest shelly fossils, a curious Early Cambrian genus named after the pioneering trilobite specialist John W. Salter. First described in 1861 from the shores of Labrador (where I have collected thousands of the little conical shells around some of the earliest metazoan reefs), its affinities can only be guessed: is it a worm, a coral, a mollusk?

Ordovician "sea beetle." Guaranteed an excellent fossil record by their calcite carapaces, trilobites are the characteristic creatures of the Early Paleozoic. (*Ceraurus pleurexanthemus*, from Ontario.)

CHIP CLARK / NMNH / SMITHSONIAN INSTITUTION

Coincidence, circumstance, and chance, and their effects on the global gene pool through time, are pervasive themes articulated throughout the book. At the personal level, Fortey explores how one chooses a career path, who happens to win the prizes and scholarships, and who loses out to disappear from sight. In the fossil record we learn about the luck of the gene draw, evolution through the trials of mass extinctions, the consequences of changing climates, continental drift, and cosmic impacts.

The book has many strengths. Fortey lyrically raises fossils from the dead, re-creating vibrant, vivid organisms that absorb light, breathe, eat, function, and interact with their ecosystems. Read his descriptions of the Middle Cambrian Burgess Shale from Canada ("on the dark shales there was a fishmonger's slabful of arthropods"), a Carboniferous rainforest ("the air is so humid that the moisture congeals upon your shoulders"), and the Eocene Messel Grube from Germany ("imagine a delicate bat, *Palaeochiropteryx*, as fragile as a paper kite, with every bone laid out upon a dark slab, as if it had been waiting its turn as an extra in a Dracula movie"). The author presents bites of life's story sequentially, from oldest to newest, as if to suggest (probably rightly so) that the past is the key to understanding the present and the future. He moves continents about like cardboard cut-outs to explain migration paths of continental tetrapods and plants. He lucidly spells out the "rules of the evolutionary game" (which organisms needed to follow to succeed, compete, and survive over millenia), and how these are displayed in the fossil record. Fortey provides a bird's eye view of the science of paleontology, and an insider's perspective of the "psycho-cultural" shenanigans that often come with the paleo-priesthood: the cladist cult, the mass extinction dichotomy of catastrophists and uniformitarians, the taxonomic schism of splitters and lumpers, the heretic leaders, and the hermits who wait in isolation to reach

---

The author is at the Department of Earth Sciences, Laurentian University, Sudbury, Ontario, Canada P3E 2C6. E-mail: pcopper@nickel.laurentian.ca

The New York Times MAY 10 1998

Circ: 1,767,836

Scientist's Plan: Map All DNA Within 3 Years

By NICHOLAS WADE A1.

A pioneer in genetic sequencing and a private company are joining forces with the aim of deciphering the entire DNA, or genome, of humans within three years, far faster and cheaper than the Federal Government is planning.

If successful, the venture would outstrip and to some extent make redundant the Government's $3 billion program to sequence the human genome by 2005.

Despite a host of new questions, the charting of the full human genome would offer enormous medical and scientific benefits.

The principals have high credibility in the world of genome sequencing. They are Dr. J. Craig Venter, president of the nonprofit Institute for Genomic Sciences in Rockville, Md., and Michael W. Hunkapiller, president and technical maestro of the Applied Biosystems division of the Perkin-Elmer Corporation of Norwalk, Conn.

The director of the Federal human genome project at the National Institutes of Health, Dr. Francis Collins, first heard of the new company's plan on Friday, as did the director of the N.I.H., Dr Harold Varmus. Both said that the plan, if successful, would enable them to reach a desired goal sooner. Dr. Collins said he planned to integrate his program with the new company's initiative. The Government would adjust by focusing on the many projects that are needed to interpret the human DNA sequence, such as sequencing the genomes of mice and other animals.

Both Dr. Varmus and Dr. Collins expressed confidence that they could persuade Congress to accept the need for this change in focus, noting that the sequencing of mouse and other genomes has always been included as a necessary part of the human genome project.

Mr. Hunkapiller's unit is a principal manufacturer of the machines used to sequence DNA, or determine the order of chemical units. The venture will be financed by Perkin-Elmer, a longtime scientific instrument maker that has recently branched into the genome field under the leadership of its new chief executive, Tony L. White.

A plan to form a new company for the venture was approved by Perkin-Elmer's board on Friday afternoon. The project could have wide ramifications for industry, academia and the public because it would make possible almost overnight many developments that had been expected to unfold over the next decade.

One such development is individualized medicine, the tailoring of drugs and other treatments to patients depending on specific variations in their DNA sequence. The wide availability of individual DNA sequences would raise more urgently the longstanding but unresolved issues of privacy and control of genetic information.

The possible possession or control of the entire human genome by a single private company could also become an issue of public concern.

The new venture was conceived only a few months ago. Mr. Hunkapiller believed that a new generation of sequencing machines coming on line would be so fast that the whole human genome could be completed far sooner and 10 times more cheaply than envisaged by the National Institutes of Health.

He approached Dr. Venter, who had developed the idea for a new sequencing strategy but lacked the means to execute it. The two men concluded in January that it would be possible to sequence the three billion letters of human DNA within three years, at a cost of $150 million to $200 million.

The $3-billion Federal program, by contrast, is now at the halfway point of its 15-year course, and only 3 percent of the genome has been sequenced. The strategy has been to divide the task and assign parts to various universities. Although the program has had many successes in pioneering a daunting task, serious doubts have emerged as to whether the universities can meet the target date of 2005.

The human genome contains all the instructions — some 60,000 or so genes — needed to design and operate the human organism. Deciphering the script in which the instructions are written — the chemical units of DNA — would yield a trove of knowledge about human physiology and disease, as well as the power, in principle, to correct the errors in DNA programming that cause genetic disease. The genome, once deciphered, is likely to be seen as the foundation of human biology, and hence is the object of intense scientific and commercial interest.

The proposal to substantially complete the human genome in three years would seem extreme hubris coming from almost anyone but Dr. Venter. But other experts deemed his approach technically feasible.

"It's not impossible at all that he could succeed," said Dr. William A. Haseltine, chief executive of Human Genome Sciences of Rockville, Md. "He has demonstrated a fine track record of innovation and organization."

Dr. Haseltine's company was for several years in uneasy partnership with Dr. Venter's institute.

If successful, the new venture seems likely to impose adjustments on all the others involved in genome research, and to offer new opportunities. Congress, for instance, might ask why it should continue to finance the human genome project through the National Institutes of Health and the Department of Energy if the new company is going to finish first.

The sponsors of the new venture insist that there will be more work for the human genome project participants to do, not less, because obtaining the DNA sequence is only the first step toward understanding what the genetic instructions mean and how they operate.

A new private venture has lofty goals but also much credibility.

CONTINUED

The New York Times MAY 10 1998
Circ: 1,767,836

"There is a strong case for Congress to increase funding for this work," said Mr. White of Perkin-Elmer. "The post-genomic world will be much more exciting."

With the new company, Perkin-Elmer would seem for the first time to be stepping into direct competition with the customers who buy its sequencing machines and other genome-analysis equipment. Mr. White, however, has no evident ambitions to become the Bill Gates of the genome world.

"We are anxious to talk to anyone who might feel threatened by this to make very sure that we are doing something compatible," Mr. White said.

Even Dr. Venter, who is known for his direct approach, said, "We are trying to do this not with an in-your-face kind of attitude." He added that he intended to work closely with the National Institutes of Health.

Dr. Venter forecast that the possession of the human genome sequence would stimulate new directions in medicine and biology, just as his sequencing of the first bacterial genome has led to a wave of other microbes being spun through sequencing machines. He said he intended to build a network of collaborators around the world to work on human genetic diseases.

Dr. Venter and his new colleagues plan not just to sequence the human genome but to construct a "definitive" data base that will integrate medical and other information with the basic DNA sequence. An important component of the new data base will be human polymorphisms, the geneticists' term for commonly found variations in DNA. Though all people and ethnic groups are thought to have an overwhelmingly similar sequence of DNA letters in their genome, there are many minor variations at certain sites on the genome, and these variations make each individual unique.

The new company's data base seems likely to rival or supersede Genbank, the data bank operated by the National Institutes of Health.

Having so much information in the control of one company is also likely to be a matter of some public concern.

"The question is, can the moral and legal questions be addressed if the largest scientific revolution of the next century is going to be done under private auspices?" said Dr. Arthur Caplan, an ethicist at the University of Pennsylvania with whom Dr. Venter has discussed the new company's goals.

The issues of genetic counseling and insurance have been around for some time, Dr. Caplan noted, but the new company's plans "accentuate the need to improve statutes governing the control of genetic information."

Perkin-Elmer intends to be sparing in laying claim to intellectual property rights over the genome, believing the company will create more demand for its machines if it allows its sequences to be widely accessible. Mr. White said his company had a track record of liberally licensing its inventions so as to improve the chances of their becoming the industry standard.

Whether the new company could gain a significant lock on the human genome in terms of patents is not at all clear. Human Genome Sciences, for example, has already obtained the full-length sequence of 80 percent of human genes, Dr. Haseltine said, and has presumably filed patent applications. The new company may therefore find that others have beaten it to the treasure trove.

Even though many have now been sequenced, genes constitute only 3 percent of the total genome. Dr. Haseltine suggested that the long regions of DNA in between the genes were like cosmology, fascinating to know about but of little commercial interest.

The new company will be 80 percent owned by Perkin-Elmer, with Dr. Venter and others owning the balance. Dr. Venter said he would resign as president of the Institute for Genomic Sciences, his place being taken by Dr. Claire Fraser, his wife.

THE WALL STREET JOURNAL. MAY 11 1998
Circ: 1,852,863

Perkin-Elmer Jumps Into Race to Decode Genes

By BILL RICHARDS BV
Staff Reporter of THE WALL STREET JOURNAL

Scientific-instrument maker Perkin-Elmer Corp. said it will join one of the nation's leading genetic researchers in a bold venture to speed up the decoding of human genes.

Perkin-Elmer, a Norwalk, Conn., company that recently moved into the genetic-sequencing field, said Saturday it signed letters of intent with J. Craig Venter and Dr. Venter's Institute for Genomic Research to form the project. They said they expect state-of-the-art sequencers from Perkin-Elmer's Applied Biosystems Division to give Dr. Venter's new project greater genetic-sequencing capacity than the entire current world genetic-sequencing output.

The announcement brings a new competitor to a race already being run by a host of companies, including Incyte Pharmaceuticals and Human Genome Sciences, with which Dr. Venter was affiliated. Researchers are continually improving the speed and accuracy of decoding techniques, and it remains to be seen whether the new project represents a major advance or simply an incremental step, analysts say.

Sequencing the human genome — the sum of DNA, which contains the inherited instructions for development — is the process of identifying the precise order of the genetic letters that make up DNA. With this sequence in hand, scientists expect to be able to more easily identify the estimated 50,000 or so genes that make up the entire genetic map. Scientists hope to pinpoint all the genes sometime around the year 2010, but it will still take years after that to figure out what the genes actually do.

The stepped-up capability, the project's leaders have told federal officials, could cut as much as three or four years off the complete-decoding timetable for the human genome. The National Institutes of Health's human-genome project has sequenced only about 3% of the three billion base pairs of DNA that make up the human genome.

"This will help us to get to our goal a little sooner, and that is good news," said Dr. Francis Collins, director of the NIH's National Genetic Research Institute, which is conducting the human-genome project.

But Dr. Collins and NIH Director Dr. Harold Varmus said yesterday that researchers at the dozen genome centers now working on the federal project still will have plenty to do. "If the complete genome is like an instruction book, what Dr. Venter's group will have when they are done would be like a group of paragraphs that still need to be tied together," said Dr. Collins.

Drs. Collins and Varmus said they only learned of the new venture at a briefing on Friday. They said the project's senior officials assured them that whatever information is developed will remain in the public domain. For example, drug companies working on developing new genetically engineered pharmaceuticals would be able to go to Dr. Venter's group and license information for a fee.

In New York Stock Exchange composite trading Friday, before the news, Perkin-Elmer closed at $68.50, up 43.75 cents.

Some researchers have voiced concern that the first private company to decode the human genome would be able to completely control future genetic engineering, as software giant Microsoft Corp. has been able to control the development of computer software. "We were given assurances they don't plan to lock it up," said Dr. Collins. The new company said it "plans to make sequencing data publicly available to ensure that as many researchers as possible are examining it."

While there have been rumors in the scientific community that a private company might step up to the challenge of deciphering the entire human genome, Perkin-Elmer's venture is the first to take that step. The company said yesterday that it has developed "a breakthrough DNA-analysis technology" that will vastly speed up the sequencing process. Perkin-Elmer said its new analyzers will cost about $300,000 each and will be ready for the commercial market early next year.

The NIH's Dr. Varmus called the company's technological advance "a stepping stone" to hastening the decoding of the human genome. "They appear to have pushed technology to the next notch," Dr. Collins added.

Dr. Venter's participation in the new sequencing company gives it unusual legitimacy in a field where optimism has sometimes outstripped reality. In the past few years, Dr. Venter and his Rockville, Md., institute have pioneered methods for quickly deciphering the entire genetic sequence of bacteria. The institute recently identified the genetic sequences for microbes that cause Lyme disease, syphilis and stomach ulcers.

Under the agreement, Perkin-Elmer will own 80% of the new company, to be based in Rockville.

The New York Times

Circ: 1,187,950

MAY 12 1998

Beyond Sequencing of Human DNA

By NICHOLAS WADE C3

THE sequencing of the human genome, a historic goal in biomedical research, was snatched away last Friday from its Government sponsor, the National Institutes of Health, by a private venture that says it can get the job done faster. Now Government officials are scrambling to adjust to the stunning turn of events, saying that the task of interpreting the genome may begin much sooner now, and that there is every reason for Congress to continue to fund the project.

Having the human DNA sequence in hand much earlier than anticipated will significantly accelerate the pace of biomedical research. "People will sign on to the concept that genome sequences are the underpinning of biology," said Dr. Richard Roberts, a Nobel prize winner who is the research director of New England Biolabs. "I think we are entering the most exciting era of biology. Finally we might understand what life is and how it works. The genome is just a start."

Adjusting to a bold new entry in the genome race.

The takeover of the human genome project is a venture of unusual audacity. Almost equally remarkable is that other genome experts seem to accept with little reservation that the abductors have a reasonable chance of making good on their claim to substantially complete the human genome, starting from scratch, in three years. The National Institutes of Health had planned to complete the sequence by the year 2005, after a 15-year program costing $3 billion.

The new venture will be financed by Perkin-Elmer, the scientific instrument maker, at an estimated cost of only $200 million. The idea was conceived by Michael W. Hunkapiller, head of Perkin-Elmer's Applied Biosystems division. "I won't say Mike is a genius because he'd hit me up for a raise," Tony L. White, the chief executive of Perkin-Elmer, said last week. An aide added, "Let's just say he is smart."

Dr. Hunkapiller is one of the co-inventors, along with Dr. Leroy Hood of the University of Washington, of the DNA sequencing machines that determine the order of the chemical units in the genetic material. His division recently developed a new model of their standard sequencing machine, one that is more highly automated and allows the machines to work round the clock with very little attendance. Dr. Hunkapiller realized the new machines were so much more efficient than their predecessors that a roomful of 200 or so might be able to complete the whole human genome in just a few years.

The human genome, with 3 billion units of DNA altogether, is distributed over 23 chromosomes, each of which is a single DNA molecule about 100 million units long. Dr. Hunkapiller's machines can determine the order of units in fragments of DNA, which are about 500 units in length. Some 60 million of these overlapping, 500-unit pieces of DNA must then be reassembled to give the sequence of the full-length chromosomes from which they are derived.

The reassembly process is far from straightforward, and Dr. Hunkapiller turned to Dr. J. Craig Venter, a leading DNA sequencer who heads the Institute for Genomic Research in Rockville, Md. He invited Dr. Venter to a meeting and told him he thought it might be possible to sequence the whole genome. "Craig said, 'You've got to be crazy,' " Dr. Hunkapiller said. "We spent a few days working through the math and came away thinking maybe it's doable. They went back and redid the calculations and so did we."

The idea of a single organization cracking the genome in a single procedure, known as a shotgun experiment, is extremely bold. Under the approach adopted by the National Institutes of Health, half a dozen university laboratories are working on the sequence, each tackling a different chromosome.

Dr. Francis Collins, the N.I.H. director of the human genome project, is proud of their progress, noting that 4 percent of the genome has already been sequenced, whereas the initial plan called for only 1 percent to be completed by this stage. But some scientists in the biotechnology industry say N.I.H.'s management of this industrial-scale project has been flawed from the start.

"There have been serious problems of organization and management both at the Department of Energy and at N.I.H.," together with internal dissension among the senior scientists involved, said Dr. William A. Haseltine, chief of Human Genome Sciences, a genome sequencing company in Rockville, Md.

That issue will be moot if the sequencing of human DNA is assumed by the new private venture. However, it is hard to see how the new venture could have started without the substantial groundwork laid by N.I.H. and by the university programs it funded, particularly the team at Washington University at St. Louis, led by Dr. Robert Waterston.

Recognizing the credibility of the new venture by Dr. Venter and Perkin-Elmer, N.I.H. officials are preparing to persuade Congress to continue funding the genome project but to switch the focus from getting the sequence to the enormous tasking of interpreting it. Dr. Venter plans to enter his findings in a public database.

One essential aid to understanding the human genome is to sequence the surprisingly similar genome of the mouse. Though all biologists recognise the need for such a project, it may not be immediately clear to members of Congress that, having forfeited the grand prize of human genome sequence, they should now be equally happy with the glory of paying for similar research on mice.

The new venture accentuates the emerging importance of genomics as

CONTINUED

The New York Times

MAY 12 1998

Circ: 1,187,950

the central framework of biology and medicine. "There is a real treasure trove to be found in the total genome and its evolutionary history, particularly as other genomes, those of chimpanzees, new and old world monkeys and mice, become sequenced," said Dr. Haseltine. "Once that picture is put together we'll have a very good idea of our evolutionary history."

The Washington Post

Ci:c: 852,262

MAY 12 1998

Private Firm Aims to Beat Government To Gene Map

y Justin Gillis
nd Rick Weiss
ashington Post Staff Writers

A1

Scientists yesterday said they ould form a new company in Rock- lle that aims to unravel the entire uman genetic code by the year)01, four years sooner than the deral government expects to com- ete a similar project.

The privately funded enterprise, hich backers said could be complet- d for perhaps one-tenth the cost of e government program, raised im- ediate questions about the rele- nce and future of the $3 billion, -year federal effort. It also raised esh concerns about the prospect of e human genetic code being expro- iated by entrepreneurs who plan to tent and sell access to the most edically valuable parts.

Some biotechnology experts not volved in the new company raved out the venture, saying it promises generate enormous amounts of netic data that may quickly be nslated into better diagnostic tests d treatments for diseases.

But other experts expressed skepticism that the company could hieve its ambitious goals, saying e new technology remains unprov- n and the novel analytical approach be used may generate less useful information than other methods.

Federal officials said the accelerating government effort to find and decode all 60,000 or more genes in the human body would remain on its current course for the next 12 to 18 months, by which time it will be clearer whether the project should change its approach to accommodate the new players in the field.

"It would be vastly premature to go out and . . . change the plan of our genome centers," said Francis Collins, head of the National Human Genome Research Institute, the branch of the National Institutes of Health that co-directs the federal effort with the Department of Energy.

The new company—not yet named—will be led by J. Craig Venter, a pioneer in finding fast, cheap ways to decode genetic information. It will be backed by Perkin-Elmer Corp. of Norwalk, Conn., a major supplier of equipment for genetic analysis, and will depend on machines developed by Perkin-Elmer.

The new company will lease space near Shady Grove Adventist Hospital, just off Interstate 270 in Montgomery County's booming biotechnology corridor, Venter said. The new venture, which expects to go into operation early in 1999, will be 80 percent owned by Perkin-Elmer.

The company will employ between 400 and 800 people to run 230 specialized new machines—each about the size of a minibar—that will operate 24 hours a day decoding information from human genes that have been isolated from sperm and other cells, Venter said. The electric bill alone is expected to hit $5,000 a day.

Venter helped found Human Genome Sciences Inc. of Rockville, the first private company in the nation to amass large amounts of genetic data, and now heads the nonprofit Institute for Genomic Research, also in Rockville.

Several biotechnology companies, including Human Genome Sciences, are in the business of decoding genetic information and selling it to pharmaceutical companies and others who hope to profit. Most of these biotech companies claim to have decoded more than 80 percent of human genes already, although the functions of most remain a mystery.

These companies have been granted scores of patents on their genetic discoveries, raising fears among some critics that a handful of companies will control the commercialization of a vast and potentially lucrative biological resource. Those fears arose again yesterday with Venter's announcement of his new project.

"Even though they are promising public access, they control the terms and there is a history of terms being more onerous than is acceptable to most scientists," said Maynard Olson, a medical geneticist at the University of Washington.

Venter said that with the exception of perhaps 100 to 300 genetic sequences that he expects will show special commercial promise, the company will make all the genetic information available free to the world's scientists. "It would be morally wrong to hold the data hostage and keep it secret," he said.

Perkin-Elmer senior vice president Michael W. Hunkapiller said the company will make money by analyzing the genetic information and then selling the results to pharmaceutical companies. The company also plans to analyze the tiny genetic differences between individuals, as opposed to getting a "generic" genetic sequence for the average human being. That new level of information, also being sought by federal laboratories, may help drug companies customize medicines for individuals or small groups of people.

Venter's technique will differ markedly from that being used by biotech companies. Those companies use a shortcut that deliberately omits large amounts of information whose role in the body is unclear.

By contrast, Venter's project aims to unravel every bit of genetic information, regardless of whether it's suspected to be useful, and to organize the resulting database into a massive and readily

The Washington Post

CONTINUED

Circ: 852,252

MAY 12 1998

consulted blueprint of human life.

To do so, the Perkin-Elmer machines will use a controversial approach called "shotgun whole genome sequencing." Instead of focusing on large pieces of DNA, this process decodes tiny pieces that later must be assembled like interlocking pieces of a jigsaw puzzle. Because of the added difficulty of dealing with so many small pieces, the resulting picture of the human genome is likely to be peppered with more and larger holes than that produced by the federal program, Collins said.

The government considered switching to the approach that Venter will use a few years ago, Collins said, and "roundly rejected" it as too problematic. But Venter and others said recent technical improvements make the approach superior.

Executives of biotechnology companies involved in genetic research have long argued that they could do the work of the federal genome project faster and more cheaply. William Haseltine, head of Human Genome Sciences, yesterday called the government's program a "gravy train" and faulted its leaders for what he described as a failure to enlist private industry.

While expressing some doubt that Venter and Perkin-Elmer would find ways to make money on their new endeavor, he said he had little doubt they would succeed in decoding the entire human genome in three years.

"This has to feel like a bomb dropped on the head of the Human Genome Project," Haseltine said by telephone from Frankfurt. "All of a sudden somebody is going to pull a $3 billion rug out from under you? They must be deeply shocked."

The Washington Times

Circ: 66,662

MAY 17 1998

Genetic mapping triggers contest

Academics race private enterprise

By Clive Cookson
FINANCIAL TIMES

C16

LONDON — The race between academic and commercial interests to unravel the entire human genetic code took another twist Wednesday when the British-based Wellcome Trust, the world's largest charity, announced that it would spend an extra $184 million on the project over the next seven years.

The trust's commitment, on behalf of the public sector, is a challenge to the commercial genomics venture announced in the United States last weekend.

Perkin-Elmer, the scientific instrumentation company, said it would set up a new company with Craig Venter, president of the Institute for Genomic Research, "to substantially complete the sequencing of the human genome [all human DNA] within three years."

Wellcome said in a statement Wednesday: "The Trust is concerned that commercial entities might file opportunistic patents on DNA sequences."

The trust is conducting an urgent review of the credibility and scope of gene patents. In a clear threat to Perkin-Elmer and other commercial organizations, Wellcome said it "is prepared to challenge such patents."

The Human Genome Project — a $3 billion, 15-year effort to spell out all 3 billion chemical "letters" in human DNA — was started in 1990 in the public sector, with funding mainly from the U.S. government. But during the 1990s the private sector moved in, led by Human Genome Sciences, a U.S. biotechnology company.

Now there's intense competition — not only between gene-hunting companies but also between the private and academic sectors as a whole.

The private sector says the profit motive is accelerating the medical application of genetic information, while the academics, led by the Wellcome Trust, claim that companies are delaying progress by preventing the open release of information.

The trust's new commitment will bring its total spending on the Human Genome Project to $328 million. The work is based at Wellcome's new Genome Campus in Cambridge, England, where DNA sequences are released freely on the Internet as they are produced.

In the United States, Venter plans to use ultrafast DNA sequencing machines developed by Perkin-Elmer, together with a new scientific strategy, to move ahead faster than the public-sector genome project. The new company is expected to have a research budget of about $200 million.

Although the data will be made publicly available after a delay, the company plans to build up a commercial database and to patent some genes.

Michael Morgan, who runs Wellcome's genomics program, said Venter's shotgun approach remained speculative and had not been proved to work. "At best it will give a quick and dirty version of the genome," he said.

•*Distributed by Scripps Howard*

The New York Times MAY 17 1998
Circ: 1,767,836

International Gene Project Gets Lift

Wellcome Trust Doubles Commitment to Public-Sector Effort

By NICHOLAS WADE A20

The politics of the human genome project, the plan to sequence or analyze the entire DNA of human cells, has become suddenly more complicated, on both a personal and international level.

The project, a glittering scientific prize expected to form the underpinning of biology and medicine in the next century, is a $3 billion Federal effort, bolstered with a significant British contribution, that aims to decode the three billion chemical letters of human DNA by 2005.

This program, now half way through its 15-year course, was upstaged by the announcement on May 10 that a private company would start and aim to complete the human DNA sequence in three years at a fraction of the cost.

Now the Wellcome Trust of London, the world's largest medical philanthropy, has stepped into the fray in an effort to maintain the impetus of the publicly financed program and to prevent the human genome sequence from falling under the control of a private company.

The trust said this week that it would double the money it gives to the Sanger Centre near Cambridge, England, enabling biologists there to sequence one-third of the genome, up from their previous goal of one-sixth. In addition, the trust said it stood ready to pay for half of the entire human genome, or DNA sequence.

"To leave this to a private company, which has to make money, seems to me completely and utterly stupid," said Dr. Michael J. Morgan, program director for the Wellcome Trust.

Asked if the trust was prepared to finance the sequencing of the entire human genome, Dr. Morgan said, "If we had to and if we wanted to, we could do it." The Wellcome Trust, he noted, has assets of $19 billion.

The Wellcome Trust's firm support of the existing program seems to have had a bracing effect on its American partner, the National Institutes of Health. Officials there were talking last week of how to "integrate" their program with the commercial venture, as if there were no point in the Government continuing its sequencing efforts, and of switching their program from sequencing to understanding how the genome works. But as the rival commercial venture has come under criticism from academic scientists, the officials no longer assume it is a probable fait accompli. The new company will produce only a "rough draft" of the DNA sequence, which may not meet scientific needs, Dr. Harold Varmus, director of the N.I.H., wrote in a recent letter to The New York Times.

Dr. John E. Sulston, director of the Sanger Centre, criticized Dr. J. Craig Venter, the head of the new venture, for opting out of the international collaboration among academic centers, and for his plan to leave gaps in parts of the sequence. "I really don't see this as being any great advance whatever," he said. "We are going to provide the complete archival product and not an intermediate, transitory version of it."

The Sanger Centre has sequenced a third of the human DNA now in the data banks, a larger contribution than that of any other institution.

Politics swirls about a glittering scientific prize.

The fighting words from the N.I.H. and the Wellcome Trust suggest that these two agencies are not about to fold their hands and will continue to sequence the human genome in competition with the new company. This venture, which has yet to be named, is being financed by the scientific instrument maker Perkin-Elmer, under the direction of Dr. Venter, a leading DNA sequencer and president of the Institute for Genomic Research in Rockville, Md.

Congress will presumably face the decision of whether to continue paying for N.I.H. to sequence the genome, possibly both lagging and duplicating Dr. Venter's effort, or to have the N.I.H. switch the emphasis of its program to interpreting the genome. Sequencing the genomes of much-studied laboratory animals like the mouse and the Drosophila fruitfly would be a major part of an interpretive, post-genomic program but doubtless less glamorous, in Congress's eyes, than obtaining the human genome sequence.

Dr. Venter, a scientist who prizes his independence and has seldom been averse to criticizing the scientific establishment, says his critics are reacting from emotion and an incomplete understanding of what he proposes to do. Despite the commercial basis of his new venture, he says he will attain the same accuracy — no more than one error in 10,000 units of DNA — as the academic centers.

On the issue of completeness, Dr. Venter acknowledges he will leave certain gaps in the genome sequence but he and his critics differ on the significance. Dr. Robert Waterston, a leading DNA sequencer at the University of Washington in St. Louis, said the quality of Dr. Venter's sequence will be "very significantly compromised," with the final product being similar to "an encyclopedia ripped to shreds and scattered on the floor."

Dr. Venter said he planned to leave no gaps in the genes themselves or in any important region between the genes. "These arguments and debate are over less than 100th of 1 percent of the genome," he said.

CONTINUED

CONTINUED The New York Times MAY 17 1998

Circ: 1,767,836

Dr. Venter knows that if his project succeeds, he will force a major adjustment on his academic competitors. He alternates between offering balm and salt for his rivals' wounds. He says he seeks to cooperate with other centers and will share his raw data, the chromatographic traces from the DNA sequencing machines, on request. But he also says he plans to sequence the genome of the Drosophila fruitfly, an important laboratory organism, as a trial run for the human sequence, and adds, "We are going to do the Drosophila genome in one-tenth the time of the C. elegans sequence and more accurately."

This is a jibe at Dr. Sulston and Dr. Waterston, who expect to complete the DNA sequence of the C. elegans nematode worm, another important laboratory organism, by the end of this year. This spectacular achievement will mark the first animal genome to be sequenced.

Dr. Sulston and Dr. Waterston have collaborated for many years in a friendship that began in Cambridge. They chose the worm genome as the pilot project for their assault on the human genome

They and Dr. Venter are well known as pioneers in the field of genomics, the study of an organism's full set of genes. Dr. Sulston and Dr. Waterston have been influential in setting the technical standards of the human genome project and the ethical standards for making data immediately available to other researchers. Dr. Venter has pioneered the sequencing of bacterial genomes, a flourishing new field that is likely to have a broad impact on medicine.

The New York Times

Circ: 1,187,950

MAY 21 1998

Gene-Mapping, Without Tax Money

By William A. Haseltine

A37

ROCKVILLE, Md.

Sometimes, it's smart *not* to compete. The Energy Department and the National Institutes of Health are spending $3 billion to decode the entire human genetic structure by 2005. But this effort has recently been upstaged by a new private company founded by Dr. J. Craig Venter, president of the nonprofit Institute for Genomic Research, and the Perkin-Elmer Corporation. This venture, which will spend about $200 million, promises to complete the job in a mere three years. In response, the Wellcome Trust, a British foundation, pledged to double its $185 million grant to a nonprofit laboratory for similar work.

Decoding the entire genome would surely be a glittering scientific achievement and may lead to some scientific breakthroughs. And knowing how individual genes work and how they fail is the key to discovering new ways to predict, detect, treat and cure many, if not most, diseases.

But there is a good reason that the Federal Government should end its effort: decoding the entire genome doesn't add significantly to the information we already possess.

Imagine that the genome is an encyclopedia with about three billion letters. Buried within this text are about 100,000 sentences (the genes) that tell the body what essential proteins to make.

LIFE BY DESIGN

The sentences are separated from one another by page after page of random letters — what scientists call junk DNA. To make matters even more complicated, the sentences themselves are also fragmented and interrupted by pages and pages of random letters — more junk DNA. In fact, less than 5 percent of our DNA contains real information. The other 95 percent has no genetic meaning.

How do we know this is really true? We've already decoded 3 percent of the entire genome. And this is the picture we get.

Each of the human genome projects, however, seeks to read the entire text from beginning to end — regardless of whether the information is useful. And regardless of the fact that we've already decoded the useful DNA.

About eight years ago, a new means to discover genes using computerized robots was developed. This method takes advantage of the fact that the human body is an excellent editor, that it can splice together the gene fragments to form a coherent text.

Instead of searching for relevant gene fragments within junk DNA, the new robotic method ignores the junk DNA and isolates only the body's edited text. This new method has been used to discover about 100,000 useful genes — almost a complete set. (My company has filed patents on more than 500 of these genes.) This information is now available for medical research; much of it is even on the World Wide Web.

So it makes little sense for the Federal Government to go to the trouble of decoding the junk DNA. Today's task is to discover the medical uses of each gene and to find gene-based cures for cancer, heart disease, Alzheimer's, osteoporosis and other diseases. The $3 billion of Federal money now devoted to the entire human genome should be spent instead on university-based research, initiated by individual medical investigators.

The era of government-sponsored big science, in which a few laboratories receive as much as $10 million a year to analyze mostly junk DNA, while scientists doing disease-related research beg for financing, should end.

Let private companies and charitable foundations finish the job of sequencing the human genome. National pride should come from conquest of disease, not winning a race that is not worth winning.

William A. Haseltine, a professor at Harvard Medical School from 1976 to 1993, is chief executive officer of Human Genome Sciences, which does gene research. From 1992 to 1996, his company helped finance the Institute for Genomic Research.

PETER SIBBALD

FORMIDABLE
Venter plans to finish the genetic code in three years—with Perkin-Elmer picking up the tab

BIOTECH

THE DUO JOLTING THE GENE BUSINESS

Craig Venter and Perkin-Elmer target the human genome

In late 1997, an ambitious idea occurred to technology guru Michael W. Hunkapiller of Perkin-Elmer Corp. Hunkapiller's team was developing a robotic machine that promised to decipher human genes far faster and more cheaply than any previous system. Why not use the new device, Hunkapiller wondered, to tackle one of the biggest prizes in all of biology—successfully deciphering the entire human genetic code? He brought his idea to gene sleuth extraordinaire J. Craig Venter, president of the nonprofit Institute for Genomic Research in Rockville, Md.

The result, announced on May 9, is a still unnamed company that will decipher what one "might describe as the full Monty—the entire genome," says Venter. With some 230 of the new $300,000 Perkin-Elmer machines running around the clock, Venter and colleague Mark Adams figure they can break the 3 billion individual units of human DNA—the genome—into pieces and decode a staggering 100 million individual units a day. They plan to finish the genetic code in three years, at a total cost of about $200 million—with Perkin-Elmer picking up the tab. That is a fraction of what the federal government is spending to complete the task—and Venter vows to finish four years sooner.

What's more, Venter and Perkin-Elmer will give away the entire human DNA sequence, just as the government plans to do. We agreed it would be morally wrong to hold the data hostage," says Venter. The gamble for Perkin-Elmer—a pioneer in gene sequencing—is that it can make money by selling information about what the sequence means, as well as finding new genes for developing medical therapies.

"BLACK EYE." The announcement sent shock waves through the red-hot field of gene-mining. This discipline, called genomics, is already populated by dozens of companies (table, page 72) and academic labs seeking to understand and profit from DNA's secrets. Companies such as Human Genome Sciences Inc. (HGS) and Incyte Pharmaceuticals Inc. have already made millions selling access to their private stashes of gene sequences. But the new company is a formidable competitor—"a 1,000-pound gorilla," says analyst Elizabeth Silverman of BancAmerica Robertson Stevens. Adds Randal W. Scott, president of Incyte in Palo Alto, Calif: "This puts a new competitor into play." And the idea that a private company can soundly beat the existing taxpayer-funded effort to the prize "is a tremendous black eye for the government," says William A. Haseltine, CEO of HGS. "They will lose the race to the genome."

But the venture also raises a host of questions. Does the massive private effort mean that the government's Human Genome Project should redirect its efforts? And will Perkin-Elmer actually be able to make money from its radically different business plan?

On the science, few are betting against Venter. "There's no question that the person who can put together an operation like this and make more headway than anyone else is Craig Venter," says Stanford University biochemist and Nobel laureate Paul Berg. Back in the mid-1990s, Venter pioneered a "shotgun" approach to deciphering entire genomes. The idea was to chop the DNA of an organism into pieces, decipher each of them, and then use computers to compare and assemble them in the right order. Using the technique, Venter astounded the scientific world by decoding the first complete genetic sequence of a living organism—a bacterium called *Haemophilus influenzae*.

Perkin-Elmer's new machines will speed up the process. Its Applied Biosystems Division sold $650 million worth of DNA sequencers and related instruments and services in fiscal 1997. The new tool, available next year, "is an evolution of our current system," says Hunkapiller. Its improved sensitivity and automation will dramatically boost productivity.

WHO'S WHO IN GENES

Craig Venter's new venture is entering a crowded field. Here are some key players that want to unlock the secrets of genes.

GENSET Under top French molecular biologist Daniel Cohen, it offers genetic information to help drugmakers tailor drugs to individuals.

HUMAN GENOME SCIENCES A pioneer in this field, HGS has built a vast database of genes, some of which it is using to create novel drugs.

HYSEQ has developed technology for rapid sequencing. Collaborates with Perkin-Elmer, and is using its own tools in drug discovery.

INCYTE Owns a huge database of genes and gene fragments, and sells both its sequences and related biological information. Collaborates with microchip firms to do rapid gene analysis.

MERCK Funded a large gene-hunting project at Washington University, St. Louis. All of its findings have been deposited in public databases.

MYRIAD Discovered the breast cancer gene by studying the genes of affected families. Now searching for more genes and developing diagnostic tests.

AXYS Finding genes for diseases such as asthma, then searching for drugs to tackle the diseases.

DATA FLOOD. Venter is hinting that the government's genome project should shift its focus to, perhaps, sequencing the DNA of animals instead of people. That's not likely. Dr. Francis Collins, head of the National Institutes of Health's genome center, wants more proof that the new company will live up to its promises before he alters his course. And even if Venter succeeds, making sense of the flood of information won't be easy. Only about 3% of human genetic material is actual genes. Some of the remaining 97% of the DNA turns genes on and off, and scientists think that much of the rest is meaningless junk. Part of Venter's job will be to figure out what's what, and that could be tough. "The genes jump right out at you in microbial sequences," says Richard K. Wilson of Washington University's gene-sequencing center. "In humans, it's much more difficult."

Many are confident of Venter's scientific claims, but the business end of this venture is another story. Perkin-Elmer faces an uphill battle convincing the biotech world that this is a moneymaking idea. "What they're describing is not a commercial venture," says Incyte's Scott. "It's really Craig Venter going after the Nobel prize for sequencing the genome." HGS's Haseltine is also skeptical. "The human genome project has never been a commercial venture," he says. "This is more in the tradition of the Mellons and Carnegies"—funding a project that promises mainly to push back the bounds of knowledge.

Perkin-Elmer execs insist that their proposal has been misunderstood. "People still don't see how, if we give away the data, we will make money," sighs CEO Tony L. White, as he patiently explains the plan. Stanford's Berg says that "the big game is how to make use of the information," and that's the information White plans to sell. Rival Incyte is already an old hand at this. In fact, one of its products is a repackaging of publicly available data in more usable form, says analyst Mike G. King of Vector Securities International. Haseltine wonders how Perkin-Elmer can do this "better than the rest of the world combined."

Venter and Perkin-Elmer execs retort that the new company will have enough experience and smarts to be a leader in this toughly competitive field. They envision signing up hundreds of thousands of subscribers—both companies and academics—for a database that offers such vital information as which sequences are genes, what the genes do, and how genes can vary from person to person. Such variations, called "polymorphisms," determine whether individuals are susceptible to certain diseases or how well drugs will work. Doctors and pharmaceutical companies can use the information to better diagnose and treat people based on their genetic variations. And companies such as Affymetrix Inc. will benefit, analysts predict. Affymetrix makes gene chips, which can almost instantly spot the presence of thousands of different genes or gene variations.

DRUG DEVELOPMENT. Perkin-Elmer should also benefit. The $1.4 billion company has moved aggressively to acquire companies and new technology, transforming Perkin-Elmer from an instrument maker to one that provides services and information as well. Since White took over in 1995, the company has acquired Tropix, a leader in screening drug candidates, and GenScope, developer of gene expression technology, and forged partnerships with other players. For instance, it teamed up last June with gene-chip developer Hyseq Inc., whose products can be used to search for gene variations.

Venter's and Perkin-Elmer's venture may also profit from new genes that Venter finds. The main current approach for finding genes involves fishing out those that are actually turned on in cells. Venter argues that this tack, which, ironically, he pioneered, misses some of the genome's real gold. That's because some genes may turn on too rarely to be discovered. He estimates that by sequencing the entire genome, he'll find 10,000 to 20,000 new genes. Many will be genes for vital signaling pathways in the body and brain—ideal candidates or targets for drugs. As a result, "these genes will have tremendous value on their own," he says. He expects the new company to pluck out a few hundred of the most promising to patent and use for drug development.

The risks, of course, are high. Haseltine and others think the new company may very well succeed at deciphering the entire human genome. Making money, however, will be harder. Venter knows that, but thinks he'll prove the skeptics wrong within a year. By then, he and his supporters believe, the new tools will prove their worth, and vindicate Venter's hunches once again.

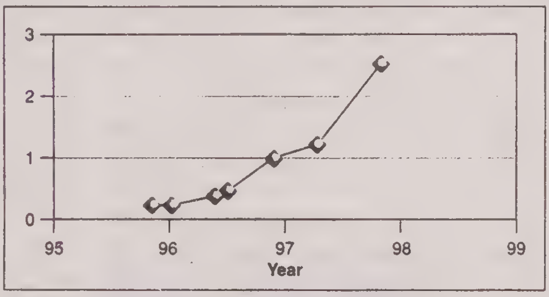
By John Carey in Washington

POLICY: BIOMEDICINE

An Independent Perspective on the Human Genome Project

Steven E. Koonin

The U.S. Human Genome Project (HGP) is a joint effort of the Department of Energy and the National Institutes of Health, formally initiated in 1990. Its stated goal is "... to characterize all the human genetic material—the genome—by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining the complete sequence ... to discover all of the more than 50,000 human genes and render them accessible for further biological study." The original 5-year plan was updated and modified in 1993 (*1*, *2*).



Percentage of the human genome sequenced to date. Almost 3% of the genome has been sequenced in contiguous stretches longer than 10 kb and is now deposited in publicly accessible databases. Compiled by J. Roach, as described in http://weber.u.washington.edu/~roach/human_genome_progress2.html.

DOE's Office of Biological and Environmental Sciences recently chartered the JASON group to review the DOE component of the HGP. This group, mainly consisting of physical and information scientists, was asked to consider three areas: technology, quality assurance and quality control, and informatics. This article summarizes the group's findings and recommendations (*3*).

Technology. The present state of the art for determining the sequence of DNA is defined by Sanger sequencing, in which DNA fragments are labeled by fluorescent dyes and separated according to length with polyacrylamide gel electrophoresis (PAGE) (*4*). The base at the end of each fragment can then be visualized and identified by the dye with which it reacts. Although more than 95% of the genome remains to be sequenced, roughly 55 megabases (Mb) have been completed in the past year (see the figure). The world's large-scale sequencing capacity (not all of which is applied to the human genome) is estimated to be roughly 100 Mb per year. It is sobering to contemplate that an average production of 400 Mb will be required each year to complete the human sequence by the target date of 2005.

The present technology has only a limited read-length capability (the number of contiguous bases that can be identified from each fragment); the best current practice can read 700 to 800 bases, with perhaps 1000 bases as the ultimate limit. Because the DNA segments of interest are much longer than this [40 kilobases (kb) for a cosmid clone; 100 kb or more for a bacterial artificial chromosome or a gene], the present technology requires that long lengths of DNA be cut into overlapping short segments (~1 kb in length) that can be sequenced directly. The sequences from these shorter pieces must then be assembled into the final sequence. Up to 50% of the effort at some sequence centers goes into this final assembly and finishing of the sequence. The ability to read longer fragments would step up the pace and quality of sequencing.

Apart from the various genome projects, however, there is little pressure to achieve longer read lengths. The 500 to 700 base lengths read by the current technology are well suited to many scientific needs, including pharmaceutical searches, studies of some polymorphisms, and studies of some genetic diseases.

Other drawbacks of the present technology include the time- and labor-intensive nature of gel preparation and running, as well as the comparatively large amounts of sample required, which also increases the cost of reagents and necessitates extra amplification steps.

Thus, the present sequencing technology leaves much to be desired and must be supplanted in the long term if the potential for genomic science is to be fully realized. Promising methods that could be cheaper and faster than PAGE include single-molecule sequencing, mass spectrometric methods, hybridization arrays, and microfluidic capabilities. None of these is sufficiently mature, however, to be a candidate for near-term major scale-up. It is therefore important to support research aimed at improving the present method. Advances in hardware development could, for example, increase the lateral scan resolution of the machine so that more lanes of a gel can be analyzed. The genome community should unify its efforts to enhance the performance of present-day instruments.

Better software will improve the lane tracking, base identification, assembly, and finishing processes. Many of the problems of base identification also occur in the demodulation of signals in communication and magnetic recording systems, and some of the existing literature in these areas should be used by the HGP. The ability to correctly assemble a final sequence without manual editing would markedly speed up the process. It would also be helpful to develop a common set of finishing rules.

Because sequencing technology should (and is likely to) evolve rapidly, the large-scale sequencing centers must be flexible enough to incorporate new technologies. There is a great need to support the development of non-PAGE–based sequencing that goes beyond the current goals of a faster version of PAGE. The funding for such advanced technology is a small fraction of the total HGP but should be increased by approximately 50%.

Quality assurance and quality control. DOE and NIH are recognizing that the HGP must make data accuracy and data quality integral to its execution. A high-quality database can provide useful, densely spaced markers across the genome and enable large-scale statistical studies. A quantitative understanding of data quality across the whole genome sequence is thus almost as important as the sequence itself. Among the top-level steps that should be taken are allocating resources specifically for quality issues and establishing a separate research program for quality assurance and control (perhaps a group at each sequencing center).

The author is professor of Theoretical Physics and vice president and provost at the California Institute of Technology. He led the JASON study reported on in this article. E-mail koonin@caltech.edu

The stated accuracy goal of the HGP is one error in 10^4 bases, which is set to be less than the polymorphism rate. However, this has been a controversial issue, as genomic data of lower accuracy are still of great utility. For example, pharmaceutical companies searching for genes can use short sequences (400 bases) at an accuracy of one error per 100 bases. The debate on error rates should focus on the level of accuracy needed for each specific scientific objective or use of the genome data. The necessity of finishing sequences without gaps should be subject to the same considerations.

In the real world, accuracy requirements must be balanced against what users need, the cost, and the capability of the sequencing technology to deliver a given level of accuracy. Establishing this balance requires an open dialogue among the sequence producers, sequence users, and the funding agencies, informed by quantitative analyses and experience.

Assays should be developed that can accurately and efficiently measure sequence quality. For example, it would be appropriate to develop, distribute, and use "gold standard" DNA samples that could be used routinely by the whole sequencing community for assessing the quality of the sequence output.

Research into the origin and propagation of errors through the entire sequencing process is fully warranted. We see two useful outputs from such studies: (i) more reliable descriptions of expected error rates in final sequence data, as a companion to database entries; and (ii) "error budgets" to be assigned to different segments of mapping and sequencing processes to aid in developing the most cost-effective strategies for sequencing and other needs.

DOE and NIH should solicit and support detailed Monte Carlo computer simulation of the complete mapping and sequencing processes. The basic computing methods are straightforward: a reference segment of DNA (with all of the peculiarities of human sequence) is generated and subjected to models of all steps in the sequencing process; individual bases are randomly altered according to errors introduced at the various stages; and the final reconstructed segment or simulated database entry is compared with the input segment and errors are noted.

Results from simulations are only as good as the models used for introducing and propagating errors. For this reason, the computer models must be developed in close association with technical experts in all phases of the process being studied, so that they best reflect the real world. This exercise will stimulate new experiments to validate the error-process models and thus will lead to increased experimental understanding of process errors as well.

Improved software is needed to enhance the ability of database centers to check the quality of submitted sequence data before its inclusion in the database. Many of the current algorithms are highly experimental and will be improved substantially over the next 5 years. In addition, an ongoing software quality assurance program should be considered for the large community databases, with advice from commercial and academic experts on software engineering and quality control. It is appropriate for the HGP to insist on a consistent level of documentation, both in the published literature and in user manuals, of the methods and structures used in the database centers that it supports. DOE and NIH should also decide on standards for the inclusion of quality metrics for base identification and DNA assembly along with every database entry submitted.

Informatics. Genome informatics is a child of the information age, a status that brings clear advantages and new hurdles. Managing such a diverse, large-scale, rapidly moving informatics effort is a considerable challenge for both DOE and NIH. The infrastructure supporting the requisite software tools ranges from small research groups (for example, for local special-purpose databases) to large Genome Centers (for process management and robotic control systems) to community database centers (for GenBank and the Genome Database). The resources that each of these groups can put into increasing software sophistication, into ensuring ease of use, and into quality control vary widely. Thus, in informatics areas requiring new research (such as gene finding), a broad-based approach of "letting a thousand flowers bloom" is most appropriate. At the other end of the spectrum, DOE and NIH must impose community-wide standards for software consistency and quality in areas of informatics in which a large user community will be accessing major genome databases.

DOE and NIH should adhere to a bottom-up, customer approach to informatics. Part of this process would be to encourage forums, including close collaborative programs, between the users and providers of informatics tools, with the purposes of determining what tools are needed and of training researchers in the use of new methods.

To ensure that all the database centers are user-oriented and that they are providing services that are genuinely useful to the genome community, each database center should be required to establish its own "users group" (as is done by facilities as diverse as the National Science Foundation's Supercomputer Centers and NASA's Hubble Space Telescope). Further, informatics centers must be critically evaluated as to the actual use of their information and services by the community.

Data formats, software components, and nomenclature should be standardized across the community. If multiple formats exist, it would be worthwhile to invest in systems that can translate among them. Data archiving, data retrieval, and data manipulation should be modularized so that one database is not overextended, and several groups should be involved in the development effort. The community should be supporting several database efforts and promoting standardized interfaces and tools among those efforts.

Final notes. The HGP involves technology development, production sequencing, and sequence utilization. Greater coupling of these three areas can only improve the project. Technology development should be coordinated with the needs and problems of production sequencing, whereas sequence generation and informatics tools must address the needs of data users. Promotion of such coupling is an important role for the funding agencies.

The HGP presents an unprecedented set of organizational challenges for the biology community. Success will require setting objective and quantitative standards for sequencing costs (capital, labor, and operations) and sequencing output (error rate, continuity, and amount). It will also require coordinating the efforts of many laboratories of varying sizes supported by multiple funding sources in the United States and abroad.

A number of diverse scientific fields have successfully adapted to a "big science" mode of operation (nuclear and particle physics, space and planetary science, astronomy, and oceanography are among the prominent examples). Such transitions have not been easy on the scientists involved. However, in essentially all of these cases, the need to construct and allocate scarce facilities has been an important organizing factor. No such centralizing force is apparent in the genomics community, but the HGP is very much in need of the coordination it would produce.

References and Notes

1. F. Collins and D. Galas, *Science* **262**, 43 (1993).
2. The status and challenges of the HGP have been recently reviewed [L. Rowen *et al.*, *ibid.* **278**, 605 (1997)].
3. The MITRE Corporation, JASON Report JSR-97-315 (The MITRE Corporation, McLean, VA, 1997). The participants included S. Block, J. Cornwall, W. Dally, F. Dyson, N. Fortson, G. Joyce, H.J. Kimble, N. Lewis, C. Max, T. Prince, R. Schwitters, P. Weinberger, and W. H. Woodin.
4. For a basic discussion and explanation of the terminogy used, see http://www.ornl.gov/TechResources/Human_Genome/publicat/primer/intro.html

DOE/ER-0713 (Part 1)

HUMAN GENOME PROGRAM REPORT

Part 1, Overview and Progress

Date Published: November 1997

Prepared for the
U.S. Department of Energy
Office of Energy Research
Office of Biological and Environmental Research
Germantown, MD 20874-1290

Prepared by the
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830-6480
managed by
Lockheed Martin Energy Research Corporation
for the
U.S. Department of Energy
Under Contract DE-AC05-96OR22464

MAJOR EVENTS IN THE U.S. HUMAN GENOME PROJECT AND RELATED PROGRAMS

1983

LANL and LLNL begin production of DNA clone (cosmid) libraries representing single chromosomes.

1984

DOE OHER and ICPEMC cosponsor Alta, Utah, conference highlighting the growing role of recombinant DNA technologies. OTA incorporates Alta proceedings into a 1986 report acknowledging value of human genome reference sequence.

1985

* Robert Sinsheimer holds meeting on human genome sequencing at University of California, Santa Cruz.

At OHER, Charles DeLisi and David A. Smith commission the first Santa Fe conference to assess the feasibility of a Human Genome Initiative.

1986

Following the Santa Fe conference, DOE OHER announces Human Genome Initiative. With $5.3 million, pilot projects begin at DOE national laboratories to develop critical resources and technologies.

1987

DOE advisory committee, HERAC, recommends a 15-year, multidisciplinary, scientific, and technological undertaking to map and sequence the human genome. DOE designates multidisciplinary human genome centers.

* NIH NIGMS begins funding of genome projects.

1988

* Reports by OTA and NAS NRC recommend concerted genome research program.

HUGO founded by scientists to coordinate efforts internationally.

* First annual Cold Spring Harbor Laboratory meeting held on human genome mapping and sequencing.

DOE and NIH sign MOU outlining plans for cooperation on genome research.

Telomere (chromosome end) sequence having implications for aging and cancer research is identified at LANL.

1989

DNA STSs recommended to correlate diverse types of DNA clones.

DOE and NIH establish Joint ELSI Working Group.

1990

DOE and NIH present joint 5-year U.S. HGP plan to Congress. The 15-year project formally begins.

Projects begun to mark genes on chromosome maps as sites of mRNA expression.

R&D begun for efficient production of more stable, large-insert BACs.

1991

Human chromosome mapping data repository, GDB, established.

1992

* Low-resolution genetic linkage map of entire human genome published.

Guidelines for data release and resource sharing announced by DOE and NIH.

1993

International IMAGE Consortium established to coordinate efficient mapping and sequencing of gene-representing cDNAs.

DOE-NIH Joint ELSI Working Group's Task Force on Genetic Information and Insurance releases recommendations.

DOE and NIH revise 5-year goals [*Science* **262**, 43–46 (Oct. 1,1993)].

* French Généthon provides mega-YACs to the genome community.

IOM releases U.S. HGP-funded report, "Assessing Genetic Risks."

GRAIL sequence interpretation service with Internet access initiated at ORNL.

ADA	Americans with Disabilities Act
ANL	Argonne National Laboratory
BAC	bacterial artificial chromosome
cDNA	complementary deoxyribonucleic acid
CGAP	Cancer Genome Anatomy Project
DNA	deoxyribonucleic acid
DHHS	Department of Health and Human Services (NIH)
DOE	Department of Energy
EEOC	Equal Employment Opportunity Commission
ELSI	ethical, legal, and social issues
GDB	Genome Database
GRAIL	Gene Recognition and Analysis Internet Link
HERAC	Health and Environmental Research Advisory Committee
HGP	Human Genome Project, Human Genome Program
HUGO	Human Genome Organisation
ICPEMC	International Commission for Protection Against Environmental Mutagens and Carcinogens
IMAGE	Integrated Molecular Analysis of Gene Expression
IOM	Institute of Medicine (NAS)

1994

* Genetic-mapping 5-year goal achieved 1 year ahead of schedule.

Completion of second-generation DNA clone libraries representing each human chromosome by LLNL and LBNL.

Genetic Privacy Act, first U.S. HGP legislative product, proposed to regulate collection, analysis, storage, and use of DNA samples and genetic information obtained from them; endorsed by DOE-NIH Joint ELSI Working Group.

DOE Microbial Genome Program launched; spin-off of HGP.

LLNL chromosome paints commercialized.

SBH technologies from ANL commercialized.

DOE HGP Information Web site activated for public and researchers.

1995

LANL and LLNL announce high-resolution physical maps of chromosome 16 and chromosome 19, respectively.

* Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published.

* First (nonviral) whole genome sequenced (for the bacterium *Haemophilus influenzae*).

Sequence of smallest bacterium, *Mycoplasma genitalium*, completed, displaying the minimum number of genes needed for independent existence.

* EEOC guidelines extend ADA employment protection to cover discrimination based on genetic information related to illness, disease, or other conditions.

1996

Methanococcus jannaschii genome sequenced; confirms existence of third major branch of life, the Archaea.

DOE-NIH Task Force on Genetic Testing releases interim principles.

* Integrated STS-based detailed human physical map with 30,000 STSs achieves an HGP goal.

* Health Care Portability and Accountability Act prohibits use of genetic information in certain health-insurance eligibility decisions, requires DHHS to enforce health-information privacy provisions.

DOE-NIH Joint ELSI Working Group releases guidelines on informed consent for large-scale sequencing projects.

DOE and NCHGR issue guidelines on use of human subjects for large-scale sequencing projects.

* *Saccharomyces cerevisiae* (yeast) genome sequence completed by international consortium.

Sequence of the human T-cell receptor region completed.

Wellcome Trust sponsors large-scale sequencing strategy meeting in Bermuda for international coordination of human genome sequencing.

1997

DOE forms Joint Genome Institute for implementing high-throughput sequencing at DOE HGP centers.

* NIH NCHGR becomes NHGRI.

* *Escherichia coli* genome sequence completed.

Second large-scale sequencing strategy meeting held in Bermuda.

* High-resolution physical maps of chromosomes X and 7 completed.

Methanobacterium thermoautotrophicum genome sequence completed.

Archaeoglobus fulgidus genome sequence completed.

* NCI CGAP begins.

LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
MGP	Microbial Genome Project
MOU	Memorandum of Understanding
mRNA	messenger ribonucleic acid
NAS	National Academy of Sciences
NCHGR	National Center for Human Genome Research (NIH)
NCI	National Cancer Institute (NIH)
NHGRI	National Human Genome Research Institute (NIH)
NIGMS	National Institute of General Medical Sciences (NIH)
NIH	National Institutes of Health
NRC	National Research Council
OHER	Office of Health and Environmental Research
ORNL	Oak Ridge National Laboratory
OTA	Office of Technology Assessment
R&D	Research and Development
SBH	sequencing by hybridization
STS	sequence tagged site
YAC	yeast artificial chromosome

---

* DOE had limited or no involvement in this event.

Preface

More than a decade ago, the Office of Health and Environmental Research (OHER) of the U.S. Department of Energy (DOE) struck a bold course in launching its Human Genome Initiative, convinced that its mission would be well served by a comprehensive picture of the human genome. Organizers recognized that the information the project would generate—both technological and genetic—would contribute not only to a new understanding of human biology and the effects of energy technologies but also to a host of practical applications in the biotechnology industry and in the arenas of agriculture and environmental protection.

Today, the project's value appears beyond doubt as worldwide participation contributes toward the goals of determining the human genome's complete sequence by 2005 and elucidating the genome structure of several model organisms as well. This report summarizes the content and progress of the DOE Human Genome Program (HGP). Descriptive research summaries, along with information on program history, goals, management, and current research highlights, provide a comprehensive view of the DOE program.

Last year marked an early transition to the third and final phase of the U.S. Human Genome Project as pilot programs to refine large-scale sequencing strategies and resources were funded by DOE and the National Institutes of Health, the two sponsoring U.S. agencies. The human genome centers at Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory had been serving as the core of DOE multidisciplinary HGP research, which requires extensive contributions from biologists, engineers, chemists, computer scientists, and mathematicians. These team efforts were complemented by those at other DOE-supported laboratories and about 60 universities, research organizations, companies, and foreign institutions. Now, to focus DOE's considerable resources on meeting the challenges of large-scale sequencing, the sequencing efforts of the three genome centers have been integrated into the Joint Genome Institute. The institute will continue to bring together research from other DOE-supported laboratories. Work in other critical areas continues to develop the resources and technologies needed for production sequencing; computational approaches to data management and interpretation (called informatics); and an exploration of the important ethical, legal, and social issues arising from use of the generated data, particularly regarding the privacy and confidentiality of genetic information.

Insights, technologies, and infrastructure emerging from the Human Genome Project are catalyzing a biological revolution. Health-related biotechnology is already a success story—and is still far from reaching its potential. Other applications are likely to beget similar successes in coming decades; among these are several of great importance to DOE. We can look to improvements in waste control and an exciting era of environmental bioremediation, we will see new approaches to improving energy efficiency, and we can hope for dramatic strides toward meeting the fuel demands of the future.

In 1997 OHER, renamed the Office of Biological and Environmental Research (OBER), is celebrating 50 years of conducting research to exploit the boundless promise of energy technologies while exploring their consequences to the public's health and the environment. The DOE Human Genome Program and a related spin-off project, the Microbial Genome Program, are major components of the Biological and Environmental Research Program of OBER.

DOE OBER is proud of its contributions to the Human Genome Project and welcomes general or scientific inquiries concerning its genome programs. Announcements soliciting research applications appear in *Federal Register*, *Science*, *Human Genome News*, and other publications. The deadline for formal applications is generally midsummer for awards to be made the next year, and submission of preproposals in areas of potential interest is strongly encouraged. Further information may be obtained by contacting the program office or visiting the DOE home page (301/903-6488, Fax: -8521, *genome@oer.doe.gov*, URL: *http://www.er.doe.gov/production/ober/hug_top.html*).

Aristides Patrinos, Associate Director
Office of Biological and Environmental Research
U.S. Department of Energy
November 3, 1997

Contents

Introduction 1
Project Origins 1
Anticipated Benefits of Genome Research 2
Coordinated Efforts 2
DOE Genome Program 3
Five-Year Research Goals 5
Evolution of a Vision 6

Highlights of Research Progress 9
Clone Resources for Mapping, Sequencing, and Gene Hunting 9
Of Mice and Humans: The Value of Comparative Analyses 13
DNA Sequencing 14
Informatics: Data Collection and Analysis 16
Ethical, Legal, and Social Issues (ELSI) 18

Technology Transfer 21
Collaborations 21
Patenting and Licensing Highlights, FY 1994–96 22
SBIR and STTR 23
Technology Transfer Award 24
1997 R&D 100 Awards 24

Research Narratives 25
Joint Genome Institute 26
Lawrence Livermore National Laboratory Human Genome Center 27
Los Alamos National Laboratory Center for Human Genome Studies 35
Lawrence Berkeley National Laboratory Human Genome Center 41
University of Washington Genome Center 47
Genome Database 49
National Center for Genome Resources 55

Program Management 59
DOE OBER Mission 59
Human Genome Program 62

Coordination with Other Genome Programs 67
U.S. Human Genome Project: DOE and NIH 67
Other U.S. Programs 68
International Collaborations 68

Appendices 71
A: Early History, Enabling Legislation (1984–90) 73
B: DOE-NIH Sharing Guidelines (1992) 75
C: Human Subjects Guidelines (1996) 77
D: Genetics on the World Wide Web (1997) 83
E: 1996 Human Genome Research Projects (1996) 89
F: DOE BER Program (1997) 95

Glossary 101

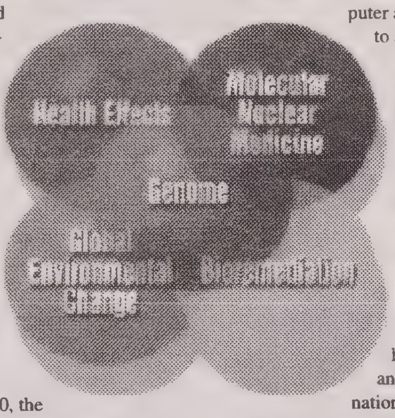
Acronym List .... Inside back cover

Introduction

genome (je nom), n, all the genetic material in the chromosomes of an organism.

Scientific and technical terms are defined in the Glossary, p. 101. More historical details and other information appear in the Appendices beginning on p. 71.

Now completing its first decade, the Human Genome Program of the U.S. Department of Energy (DOE) is the longest-running federally funded program to analyze the genetic material—the genome— that determines an individual's characteristics at the most fundamental level. Part of the Biological and Environmental Research (BER) Program sponsored by the DOE Office of Biological and Environmental Research (OBER*), the genome program is a major component of the larger U.S. Human Genome Project.



For 50 years, programs in the DOE Office of Biological and Environmental Research have crossed traditional research boundaries in seeking new solutions to energy-related biological and environmental challenges (see Appendix F, p. 95, and http://www.er.doe.gov/production/ober/ober.html*).*

Since October 1990, the project has been supported jointly by DOE and the National Institutes of Health (NIH) National Human Genome Research Institute (formerly National Center for Human Genome Research). Together, the DOE and NIH components make up the world's largest centrally coordinated biology research project ever undertaken.

The U.S. Human Genome Project is a 15-year endeavor to characterize the human genome by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining a complete sequence of the deoxyribonucleic acid (DNA) subunits. Parallel studies are being carried out on selected model organisms to facilitate interpretation of human gene function.

The ultimate goal of the U.S. project is to identify the estimated 70,000 to 100,000 human genes and render them accessible for future biological study. A complete human DNA sequence will provide physicians and researchers in many biological disciplines with an extraordinary resource: an "encyclopedia" of human biology obtainable by computer and available to all.

Obtaining the complete sequence by 2005 will require a highly coordinated and focused international effort generating advances in biological methodology; instrumentation (particularly automation); and computer-based methods for collecting, storing, managing, and analyzing the rapidly growing body of data.

Project Origins

The potential value of detailed genetic information was recognized early; until recently, however, obtaining this information was far beyond the capabilities of biomedical research. DOE OBER and its two predecessor agencies—the Atomic Energy Commission and the Energy Research and Development Administration—had long sponsored genetic research in both microbial and higher systems. These studies included explorations into population genetics; genome structure, maintenance, replication, damage, and repair; and the consequences of genetic mutations. These traditional DOE activities evolved naturally into the Human Genome Program.

---

*In 1997 the Office of Health and Environmental Research (OHER) was renamed Office of Biological and Environmental Research (OBER).

OBER's mission is described more fully in the Program Management section (p. 59) of this report.

By 1985, progress in genetic and DNA technologies led to serious discussions in the scientific community about initiating a major project to analyze the structure of the human genome. After concluding that a DNA sequence would offer the most useful approach for detecting inherited mutations, DOE in 1986 announced its Human Genome Initiative. The initiative emphasized development of resources and technologies for genome mapping, sequencing, computation, and infrastructure support that would culminate in a complete sequence of the human genome.

The National Research Council issued a report in 1988 recommending a dedicated research budget of $200 million annually for 15 years to determine the sequence of the 3 billion chemical subunits (base pairs) in the human genome and to map and identify all human genes.

To launch the nation's Human Genome Project, Congress appropriated funds to DOE and also to NIH, which had long supported research in genetics and molecular biology as an integral part of its mission to improve the health of all Americans. Other federal agencies and foundations outside the Human Genome Project also contribute to genome research, and many other countries are making important contributions through their own genome research projects.

Coordinated Efforts

In 1988 DOE and NIH signed a Memorandum of Understanding in which the agencies agreed to work together, coordinate technical research and activities, and share results. The two agencies assumed a joint systematic approach toward establishing goals to satisfy both short- and long-term project needs.

Early guidelines projected three 5-year phases, for which the first plan was presented to Congress in 1990. The 1990

Anticipated Benefits of Genome Research

Predictions of biology as "the science of the 21st century" have been made by observers as diverse as Microsoft's Bill Gates and U.S. President Bill Clinton. Already revolutionizing biology, genome research has spawned a burgeoning biotechnology industry and is providing a vital thrust to the increasing productivity and pervasiveness of the life sciences.

Technology and resources promoted by the Human Genome Project already have had profound impacts on biomedical research and promise to revolutionize biological research and clinical medicine. Increasingly detailed genome maps have aided researchers seeking genes associated with dozens of genetic conditions, including myotonic dystrophy, fragile X syndrome, neurofibromatosis types 1 and 2, a kind of inherited colon cancer, Alzheimer's disease, and familial breast cancer.

Current and potential applications of genome research will address national needs in molecular medicine, waste control and environmental cleanup, biotechnology, energy sources, and risk assessment.

Molecular Medicine

On the horizon is a new era of molecular medicine characterized less by treating symptoms and more by looking to the most fundamental causes of disease. Rapid and more specific diagnostic tests will make possible earlier treatment of countless maladies. Medical researchers also will be able to devise novel therapeutic regimens based on new classes of drugs, immunotherapy techniques, avoidance of environmental conditions that may trigger disease, and possible augmentation or even replacement of defective genes through gene therapy.

Microbial Genomes

In 1994, taking advantage of new capabilities developed by the genome project, DOE formulated the Microbial Genome Initiative to sequence the genomes of bacteria useful in the areas of energy production, environmental remediation, toxic waste reduction, and industrial processing. In the resulting Microbial Genome Project, six microbes that live under extreme conditions of temperature and pressure have been sequenced completely as plan emphasized the creation of chromosome maps, software, and automated technologies to enable sequencing.

By 1993, unexpectedly rapid progress in chromosome mapping required updating the goals [*Science* **262**, 43–46 (October 1, 1993)], which now project through 1998 (see p. 5). This plan is being revised again in anticipation of the approaching high-throughput sequencing phase of the project. Last year marked an early transition to this phase as many more genome sequencing projects were funded. The second and third phases of the project will optimize resources, refine sequencing strategies, and, finally, completely determine the sequence of all base pairs in the genome.

Another area of DOE and NIH cooperation is in exploring the ethical, legal, and social issues (ELSI) arising from increased availability of genetic data and growing genetic-testing capabilities. The two agencies established a joint working group to confront these ELSI challenges and have cosponsored joint projects and workshops.

DOE Genome Program

A general overview follows of recent progress made in the DOE Human Genome Program. Refer to the timeline (pp. ii–iii) for other achievements toward U.S. goals, including contributions made outside DOE.

Physical maps

For DOE, an early goal was to develop chromosome physical maps, which involves reconstructing the order of cloned DNA fragments to represent their specific originating chromosomes. (A set of such cloned fragments is called a library.) Critical to this effort were the libraries of individual human chromosomes

of August 1997. Structural studies are under way to learn what is unique about the proteins of these organisms—the ultimate aim being to use the microbes and their enzymes for such practical purposes as waste control and environmental cleanup.

Biotechnology

The potential for commercial development presents U.S. industry with a wealth of opportunities. Sales of biotechnology products are projected to exceed $20 billion by the year 2000. The genome project already has stimulated significant investment by large corporations and prompted the creation of new biotechnology companies hoping to capitalize on the far-reaching implications of its research.

Energy Sources

Biotechnology, fueled by insights reaped from the genome project, will play a significant role in improving the use of fossil-based resources. Increased energy demands, projected over the next 50 years, require strategies to circumvent the many problems associated with today's dominant energy technologies. Biotechnology promises to help address these needs by providing cleaner means for the bioconversion of raw materials to refined products. In addition, there is the possibility of developing entirely new biomass-based energy sources. Having the genomic sequence of the methane-producing microorganism *Methanococcus jannaschii*, for example, will enable researchers to explore the process of methanogenesis in more detail and could lead to cheaper production of fuel-grade methane.

Risk Assessment

Understanding the human genome will have an enormous impact on the ability to assess risks posed to individuals by environmental exposure to toxic agents. Scientists know that genetic differences make some people more susceptible—and others more resistant—to such agents. Far more work must be done to determine the genetic basis of such variability. This knowledge will directly address DOE's long-term mission to understand the effects of low-level exposures to radiation and other energy-related agents, especially in terms of cancer risk.

produced at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL). These libraries allowed the huge task of mapping and sequencing the entire 3 billion bases in the human genome to be broken down into 24 much smaller single-chromosome units. Availability of the libraries has enabled the participation of many laboratories worldwide. Some three generations of clone libraries with improving characteristics have been produced and widely distributed. In the DOE-supported projects, DNA clones representing chromosomes 16, 19, and 22 have been ordered (mapped) and are now providing material needed for large-scale sequencing.

Sequencing

Toward the goal of greatly increasing the speed and decreasing the cost of DNA sequencing, DOE has supported improvements in standard technologies and has pioneered support for revolutionary sequencing systems. Marked improvements have been made in reagents, enzymes, and raw data quality. Such novel approaches as sequencing by hybridization (using DNA "chips") and mass spectrometry have already found important, previously unanticipated applications outside the Human Genome Project.

Joint Genome Institute

In early 1997, the human genome centers at Lawrence Berkeley National Laboratory, LANL, and LLNL began collaborating in the Joint Genome Institute (JGI), within which high-throughput sequencing will be implemented [see p. 26 and *Human Genome News* 8(2), 1–2]. The initial JGI focus will be on sequencing areas of high biological interest on several chromosomes, including human chromosomes 5, 16, and 19. Establishment of JGI represents a major transition in the DOE Human Genome Program.

Previously, most goals were pursued by small- to medium-sized teams, with modest multisite collaborations. The JGI will house high-throughput implementations of successful technologies that will be run with increasingly stringent process- and quality-control systems.

In addition, a small component aimed at understanding how genes function in the body—a field known as functional genomics—has been established and will grow as sequencing targets are met. High-throughput functional genomics represents a new era in human biology, one which will have profound implications for solving biological problems.

Informatics

In preparation for the production-sequencing phase, many algorithms for interpreting DNA sequence have been developed, and an increasing number have become available as services over the Internet. Last year, the GRAIL (for Gene Recognition and Analysis Internet Link) and GenQuest servers, developed and maintained at Oak Ridge National Laboratory, processed an average of almost 40 million bases of sequence each month.

As technology improves and data accumulates exponentially, continued progress in the Human Genome Project will depend increasingly on the development of sophisticated computational tools and resources to manage and interpret the information. The ease with which researchers can access and use the data will provide a measure of the project's success. Critical to this success is the creation of interoperable databases and other computing and informatics tools to collect, organize, and interpret thousands of DNA clones.

For additional information on the DOE genome programs, refer to Research Highlights, p. 9; Research Narratives, p. 25; this report's *Part 2, 1996 Research Abstracts*; and the Web site (*http://www.ornl.gov/hgmis*).

Five-Year Research Goals of the U.S. Human Genome Project

October 1, 1993, to September 30, 1998 (FY 1994 through FY 1998)*

Major events in the U.S. Human Genome Project, including progress made toward these goals, are charted in a timeline on pp. ii–iii.

Genetic Mapping

- Complete the 2- to 5-cM map by 1995.
- Develop technology for rapid genotyping.
- Develop markers that are easier to use.
- Develop new mapping technologies.

Physical Mapping

- Complete a sequence tagged site (STS) map of the human genome at a resolution of 100 kb.

DNA Sequencing

- Develop efficient approaches to sequencing one- to several-megabase regions of DNA of high biological interest.
- Develop technology for high-throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
- Build up a sequencing capacity to allow sequencing at a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

Gene Identification

- Develop efficient methods for identifying genes and for placement of known genes on physical maps or sequenced DNA.

Technology Development

- Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and for meeting the needs of the Human Genome Project as a whole.

Model Organisms

- Finish an STS map of the mouse genome at a 300-kb resolution.
- Finish the sequence of the *Escherichia coli* and *Saccharomyces cerevisiae* genomes by 1998 or earlier.
- Continue sequencing *Caenorhabditis elegans* and *Drosophila melanogaster* genomes with the aim of bringing *C. elegans* to near completion by 1998.
- Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest.

Informatics

- Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.
- Consolidate, distribute, and continue to develop effective software for large-scale genome projects.
- Continue to develop tools for comparing and interpreting genome information.

Ethical, Legal, and Social Implications

- Continue to identify and define issues and develop policy options to address them.
- Develop and disseminate policy options regarding genetic testing services with potential widespread use.
- Foster greater acceptance of human genetic variation.
- Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues.

Training

- Continue to encourage training of scientists in interdisciplinary sciences related to genome research.

Technology Transfer

- Encourage and enhance technology transfer both into and out of centers of genome research.

Outreach

- Cooperate with those who would establish distribution centers for genome materials.
- Share all information and materials within 6 months of their development. This should be accomplished by submission of information to public databases or repositories, or both, where appropriate.

---

*Original 1990 goals were revised in 1993 due to rapid progress. A second revision was being developed at press time.

Evolution of a Vision: Genome Project Origins,

In an interview at a DNA sequencing conference in Hilton Head, South Carolina, David Smith, a founder and former Director of the DOE Human Genome Program, recalled the establishment of this country's first human genome project. The impressive early achievements and spin-off benefits, he noted, offer more than mere vindication for project founders. They also provide a tantalizing glimpse into the future where, he observed, "scientists will be empowered to study biology and make connections in ways undreamt of before."*

The DOE Human Genome Program began as a natural outgrowth of the agency's long-term mission to develop better technologies for measuring health effects, particularly induced mutations. As Smith explained it, "DOE had been supporting mutation studies in Japan, where no heritable mutations could be detected in the offspring of populations exposed to the atomic blasts at Hiroshima and Nagasaki. The program really grew out of a need to characterize DNA differences between parents and children more efficiently. DOE led the development of many mutation tests, and we were interested in developing even more sensitive detection methods. Mortimer Mendelsohn of Lawrence Livermore National Laboratory, a member of the International Commission for Protection Against Environmental Mutagens and Carcinogens, and I decided to hold a workshop to discuss DNA-based methods (see Human Genome Project chronology, p. ii).

"Ray White (University of Utah) organized the meeting, which took place in Alta, Utah, in December 1984. It was a small meeting but very stimulating intellectually. We concluded the obvious—that if you really wanted to use DNA-based technologies, you had to come up with more efficient ways to characterize the DNA of much larger regions of the genome. And the ultimate sensitivity would be the capability to compare the complete DNA sequences of parents and their offspring."

Project Begins

Smith recalled reaction to the first public statement that DOE was starting a program with the aim of sequencing the human genome. "I announced it at the Cold Spring Harbor meeting in May 1986, and there was a big hullabaloo." After a year-long review, a National Academy of Sciences National Research Council panel endorsed the project and the basic strategy proposed. Smith pointed out that NIH and others were also having discussions on the feasibility of sequencing the human genome. "Once NIH got interested, many more people became involved. DOE and NIH signed a Memorandum of Understanding in October 1988 to coordinate our activities aimed at characterizing the human genome." But, he observed, it wasn't all smooth sailing. The nascent project had many detractors.

> ***"Genomics has come of age, and it is opening the door to entirely new approaches to biology."***

Responding to Critics

Many scientists, prominent biologists among them, thought having the sequence would be a misuse of scarce resources. Smith, laughing now, recalls one scientist complaining, "Even if I had the sequence, I wouldn't know what to do with it." Other critics worried that the genome project would siphon shrinking research funds away from individual investigator-initiated research projects. Smith takes the opposite view. "In fact, individual investigators can do things they would never be able to do otherwise. We're beginning to see that demonstrated at this meeting. For the first time, we're finding people exploring systematic ways of looking at gene function in organisms. The genome project opens up enormous new research fields to be mined. Cottage-industry biologists won't need a lot of robots, but they will have to be computer literate to put the information all together."

The genome project also is providing enabling technologies essential to the future of the emerging biotechnology industry, catalyzing its tremendous growth. According to Smith, the technologies are capable of more than elucidating the human genome. "We're developing an infrastructure for future research. These technologies will allow us to efficiently characterize any of the organisms out there that pertain to various DOE missions, with such applications as better fuels from biomass, bioremediation, and waste control. They also will lead to a greater understanding of global cycles, such as the carbon cycle, and the identification of potential biological interventions. Look at the ocean; an amazing number of microbes are in there, but we don't know how to use them to influence cycles to control some of the harmful things that might be happening. Up to now, biotechnology has been nearly all health oriented, but applications of genome research to modern biology really go beyond health. That's one of the things motivating our program to try to develop some of these other biotechnological applications."

Responding to criticism about not researching gene function early in the project, Smith reasserted that the purpose of the Human Genome Project is to build technologies and resources that will enable researchers to learn about biology in a much

---

*The Seventh International Genome Sequencing and Analysis Conference, September 1995.

Present and Future Challenges, Far-Reaching Benefits

more efficient way. "The genome budget is devoted to very specific goals, and we make sure that projects contribute toward reaching them."

International Scope

Smith credited the international community with contributing to many project successes. "The initial planning was for a U.S. project, but the outcome, of course, is that it is truly international, and we would not be nearly as far as we are today without those contributions. Also, there's been a fair amount of money from private companies, and support from the Muscular Dystrophy Association in France and The Wellcome Trust in the United Kingdom has been extremely important."

Technology Advances

While noting enormous advances across the board, Smith cited automation progress and observed that tremendously powerful robots and automated processes are changing the way molecular biology is done. "A lot of novel technologies probably won't be useful for initial sequencing but will be very valuable for comparing sequences of different people and for polymorphism studies. One of the most gratifying recent successes is the DNA polymerase engineering project. Researchers made a fairly simple change, but it resulted in a thermosequenase that may answer a lot of problems, reduce the cost of sequencing, and give us better data."

Progress in genome research requires the use of maturing technologies in other fields. "The combination of technologies that are coming together has been fortuitous; for example, advances in informatics and data-handling technologies have had a tremendous impact on the genome project. We would be in deep trouble if they were at a less-mature stage of development. They have been an important DOE focus."

ELSI

Smith described tangible progress toward goals associated with programs on the ethical, legal, and social issues (ELSI) related to data produced by the genome project. "ELSI programs have done a lot to educate the thinkers, and this has produced a higher level of discourse in the country about these issues. DOE is spending a large fraction of its ELSI money on informing special populations who can reach others. Educating judges has been especially well received because they realize the potential impact of DNA technology on the courts."

According to Smith, more people and groups need to be involved in ELSI matters. "We have some ELSI products: the DOE-NIH Joint ELSI Working Group has an insurance task force report, and a DOE ELSI grantee has produced draft privacy legislation. Now it's time for others to come and translate ELSI efforts into policy. Perhaps the new National Bioethics Advisory Commission can do some of this."

New Model for Biological Research

Smith spoke of a changing paradigm guiding DOE-supported biology. "Some years ago, the central idea or dogma in molecular biology research was that information in DNA directs RNA, and RNA directs proteins. Today, I think there is a new paradigm to guide us: Sequence implies structure, and structure implies function. The word 'implies' in our new paradigm means there are rules," continued Smith, "but these are rules we don't understand today. With the aid of structural information, algorithms, and computers, we will be able to relate sequence to structure and eventually relate structure to function. Our effort focuses on developing the technologies and tools that will allow us to do this efficiently."

"That's how I think about what we do at DOE," he said. "We're working a lot on technology and projects aimed at human and microbial genome sequencing. For understanding sequence implications, we are making major, increasing investments in synchrotrons, synchrotron user facilities, neutron user facilities, and big nuclear magnetic resonance machines. These are all aimed at rapid structure determination." Smith explained that now we are seeing the beginnings of the biotechnology revolution implied by the sequence-to-structure-to-function paradigm. "If you really understand the relationship between sequence and function, you can begin to design sequences for particular purposes. We don't yet know that much about the world around us, but there are capabilities out there in the biological world, and if we can understand them, we can put those capabilities to use."

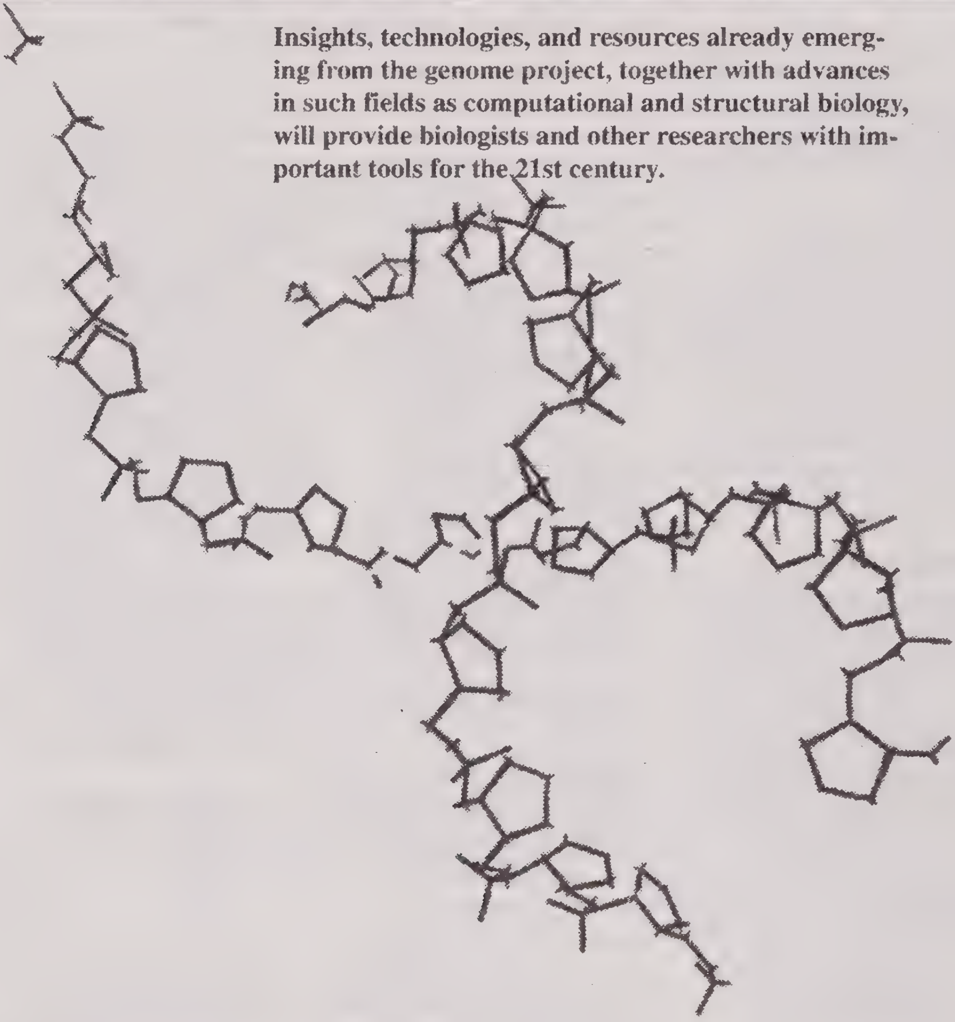
"Comparative genomics," he continued, "will teach us a tremendous amount about human evolution. The current phylogenetic tree is based on ribosomal RNA sequences, but when we have determined whole genomic sequences of different microbes, they will probably give us different ideas about relationships among archaebacteria, eukaryotes, and prokaryotes."

Feeling good about progress over the previous 5 years, Smith summed it up succinctly: "Genomics has come of age, and it is opening the door to entirely new approaches to biology."

David Smith retired at the end of January 1996. Taking responsibility for the DOE Human Genome Program is Aristides Patrinos, who is also Associate Director of the DOE Office of Biological and Environmental Research. Marvin Frazier is Director of the Health Effects and Life Sciences Research Division, which manages the Human Genome Program.

Looking to the Future

Insights, technologies, and resources already emerging from the genome project, together with advances in such fields as computational and structural biology, will provide biologists and other researchers with important tools for the 21st century.



Transitioning to large-scale sequencing

The early years of the Human Genome Program have been remarkably successful. Critical resources and infrastructures have been established, and technologies have been developed for producing several useful types of chromosomal maps. These gains are supporting the project's transition to the large-scale sequencing phase. Some highlights and trends in the U.S. Department of Energy's (DOE) Human Genome Program after FY 1993 are presented in this section.

Clone Resources for Mapping, Sequencing, and Gene Hunting

The demands of large chromosomal mapping and sequencing efforts have necessitated the development of several different types of clone collections (called libraries) carrying human DNA. Three generations of DOE-developed libraries are being distributed to research teams in the United States and abroad. In these libraries, human DNA segments of various lengths are maintained in bacterial cells.

NLGLP Libraries

The first two generations are chromosome-specific libraries carrying small inserts of human DNA (15,000 to 40,000 base pairs). As part of the National Laboratory Gene Library Project (NLGLP) begun in 1983, these libraries were prepared at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL) using DOE flow-sorting technology to separate individual chromosomes. Library availability has allowed the very difficult whole-genome tasks to be divided into 24 more manageable single-chromosome projects that could be pursued at separate research centers. Completed in 1994, NLGLP libraries have provided critical resources to genome researchers worldwide (*http://www-bio.llnl.gov/genome/html/cosmid.html*). Very high resolution chromosome maps based principally on NLGLP libraries were published in 1995 for chromosomes 16 and 19. These are described in detail in the Research Narratives section of this report (see LLNL, p. 27, and LANL, p. 35).

PACs and BACs

The third generation of clone resources supporting chromosome mapping is composed of P1 artificial chromosome (PAC) and bacterial artificial chromosome (BAC) libraries. A prototype PAC library was produced by the team of Leon Rosner (then at DuPont) many years ago, but more efficient production began with improvements introduced by the DOE-supported teams headed by Melvin Simon at Caltech (BACs) and Pieter de Jong at Roswell Park (PACs).

In contrast to cosmids, BACs and PACs provide a more uniform representation of the human genome, and the greater length of their inserts (90,000 to

DOE Genome Research Web Site
http://www.ornl.gov/hgmis/research.html

Research Narratives

Separate narratives, beginning on p. 25, contain detailed descriptions of research programs and accomplishments at these major DOE genome research facilities.

- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Lawrence Berkeley National Laboratory
- University of Washington Genome Sequencing Laboratory
- Genome Database
- National Center for Genome Resources

Research Abstracts

Descriptions of individual research projects at other institutions are given in *Part 2, 1996 Research Abstracts.*

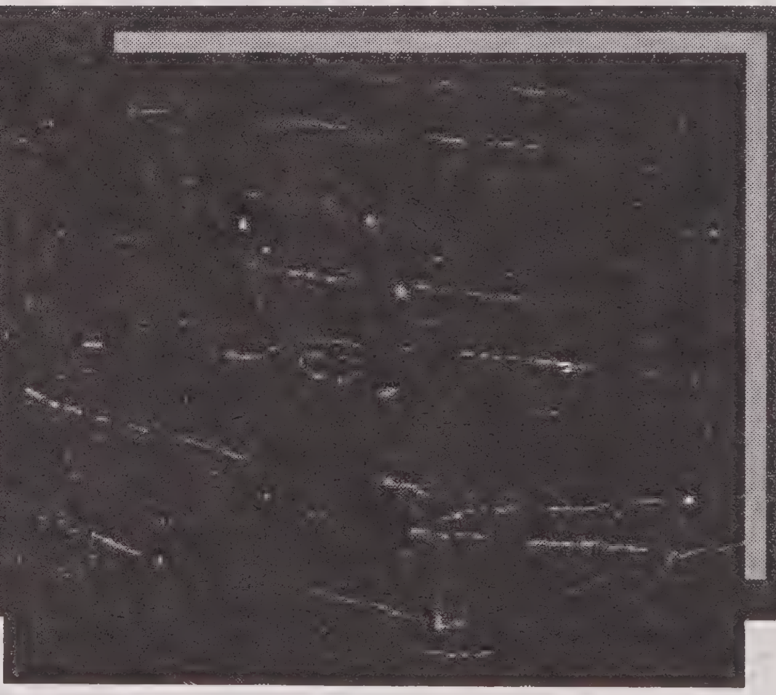
300,000 base pairs) facilitates both mapping and sequencing. Their usefulness was illustrated dramatically in 1993 when the first breast cancer–susceptibility gene (*BRCA*1) was found in a BAC clone after other types of resources had failed. The next year, with major support from NIH, de Jong's PACs contributed to the isolation of the second human breast cancer–susceptibility gene (*BRCA*2).

Mapping

The assembly of ordered, overlapping sets (contigs) of high-quality clones has long been considered an essential step toward human genome sequencing. Because the clones have been mapped to precise genomic locations, DNA sequences obtained from them can be located on the chromosomes with minimal uncertainty.

The large insert size of BACs and PACs allows researchers to visually map them on chromosomes by using fluorescence in situ hybridization (FISH) technology (see photomicrograph below). These mapped BACs and PACs represent very valuable resources for the cytogeneticist exploring chromosomal abnormalities. Two major medical genetics resources have been developed: (1) The Resource for Molecular Cytogenetics at the University of California, San Francisco, in collaboration with the Lawrence Berkeley National Laboratory (LBNL) team led by Joe Gray (*http://rmc-www.lbl.gov*) and (2) The Total Human Genome BAC-PAC Resource at Cedars-Sinai Medical Center, Los Angeles, developed by Julie Korenberg's laboratory (see map, p. 12, and Web site, *http://www.csmc.edu/genetics/korenberg/korenberg.html*).

FISH Mapping on DNA Fibers. The fluorescence microscope reveals several individual cloned DNA fibers from yeast artificial chromosomes (YACs, in blue) after molecular combing to attach and stretch the DNA molecules across a glass microscope slide. Also shown are the locations of two P1 clones, labeled green and red, mapped onto the YAC fibers using FISH. Digital imaging technology can be used to assemble physical maps of chromosomes with a resolution of about 3 to 5 kilobases. (Source: Joe Gray, University of California, San Francisco)



Coordinated Mapping and Sequencing

A simple strategy was proposed in 1996 or choosing BACs or PACs to elongate equenced regions most efficiently Nature **381**, 364–66 (1996)]. The first tep is to develop a BAC end sequence database, with each entry having the BAC clone name and the sequences of ts human insert ends. In toto, the source BACs should represent a 15- to 20-fold overage of the human genome. Then or any BAC or chromosomal region sequenced, a comparison against the database will return a list of BACs (or PACs) that overlap it. Optimal choices or the next BACs (or PACs) to be sequenced can then be made, entailing minimal everlap (and therefore minimal edundancy of sequencing).

Two pilot BAC-PAC end-sequencing projects were initiated in September of 1996 to explore feasibility, optimize echnologies, establish quality controls, and design the necessary informatics inrastructure. Particular benefits are anicipated for small laboratories that will not have to maintain large libraries of clones and can avoid preliminary contig mapping (see abstracts of Glen Evans; ulie Korenberg; Mark Adams, Leroy Hood, and Melvin Simon; and Pieter de ong in Part 2 of this report).

Updated information on BAC-PAC reources can be found on the Web (*http://www.ornl.gov/meetings/bacpac/95bac.html*). [See Appendix C: Human Subjects Guidelines, p. 77 or *http://www.ornl.gov/hgmis/archive/nchgrdoe.html* for DOE-NIH guidelines on using DNA rom human subjects for large-scale equencing.]

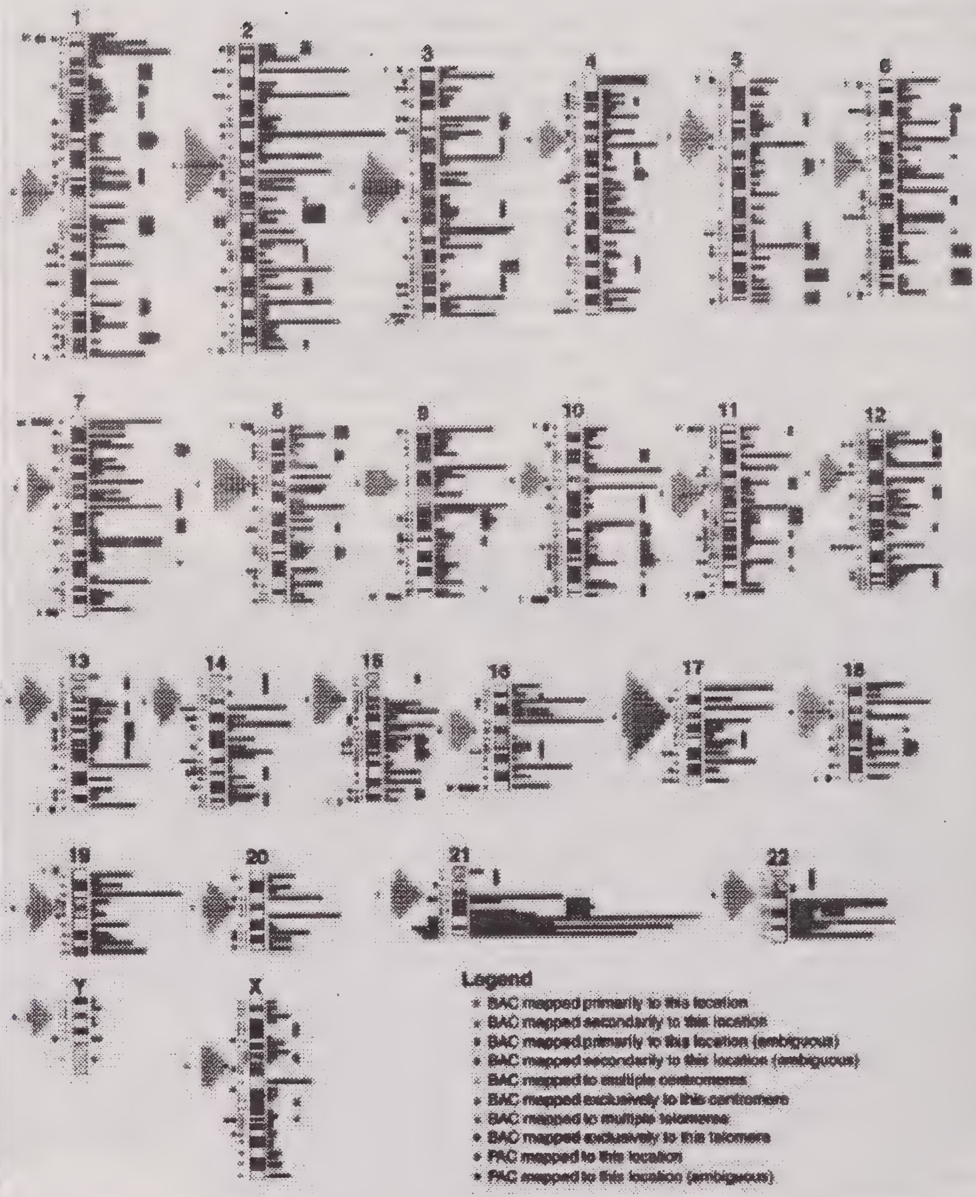
cDNA Libraries

n 1990, DOE initiated projects to enich the developing chromosome contig maps with markers for genes. Although he protein-encoding messenger RNAs re good representatives of their source genes, they are unstable and must be converted to complementary DNAs (cDNAs) for practical applications. These conversions are tricky, and artifacts are introduced easily. The team led by Bento Soares (University of Iowa) has optimized the steps and continues to produce cDNA libraries of the highest quality. At LLNL, individual cDNA clones are put into standard arrays and then distributed worldwide for characterization by the international IMAGE (for Integrated Molecular Analysis of Gene Expression) Consortium (see box, p. 13).

Initially supported under a DOE cDNA initiative, Craig Venter's team (now at The Institute for Genomic Research) greatly improved technologies for reading sequences from cDNA ends (expressed sequence tags, called ESTs). Together with complementary analysis software, ESTs were shown to be a valuable resource for categorizing cDNAs and providing the first clues to the functions of the genes from which they are derived. This fast EST approach has attracted millions of dollars in commercial investment. Mapping the cDNA onto a chromosome can identify the location of its corresponding gene. Many laboratories worldwide are contributing to the continuing task of mapping the estimated 70,000 to 100,000 human genes.

HAECs

All the previously described DNA clones are maintained in bacterial host cells. However, for unknown reasons, some regions of the human genome appear to be unclonable or unstable in bacteria. The team led by Jean-Michel Vos (University of North Carolina, Chapel Hill) has developed a human artificial episomal chromosome (HAEC) system based on the Epstein-Barr virus that may be useful for coverage of these especially difficult regions. In the broader biomedical community, HAECs also show promise for use in gene therapy.



Legend

- BAC mapped primarily to this location
- BAC mapped secondarily to this location
- BAC mapped primarily to this location (ambiguous)
- BAC mapped secondarily to this location (ambiguous)
- BAC mapped to multiple centromeres
- BAC mapped exclusively to this centromere
- BAC mapped to multiple telomeres
- BAC mapped exclusively to this telomere
- PAC mapped to this location
- PAC mapped to this location (ambiguous)

BAC-PAC Map. *The Total Human Genome BAC-PAC Resource represents an important tool for understanding the genes responsible for human development and disease (http://www.csmc.edu/genetics/korenberg/korenberg.html). The Resource, consisting of more than 5000 BAC and PAC clones, covers every human chromosome band and 25% of the entire human genome. Each color dot represents a single BAC or PAC clone mapped by FISH to a specific chromosome band represented in black and white. The clones, which are stable and useful for sequencing, have been integrated with the genetic and physical chromosome maps.* *[Source: Julie Korenberg, Cedars-Sinai Medical Center]*

Resources for Gene Discovery

Hunting for disease genes is not a specific goal of the DOE Human Genome Program. However, DOE-supported libraries sent to researchers worldwide have facilitated gene hunts by many research teams. DOE libraries have played a role in the discovery of genes for cystic fibrosis, the most common lethal inherited disease in Caucasians; Huntington's disease, a progressive lethal neurological disorder; Batten's disease, the most prevalent neurodegenerative childhood disease; two forms of dwarfism; Fanconi anemia, a rare disease characterized by skeletal abnormalities and a predisposition to cancer; myotonic dystrophy, the most common adult form of muscular dystrophy; a rare inherited form of breast cancer; and polycystic kidney disease, which affects an estimated 500,000 people in the United States at a healthcare cost of over $1 billion per year.

The team led by Fa-Ten Kao (Eleanor Roosevelt Institute) has microdissected several chromosomes and made derivative clone libraries broadly available to disease-gene hunters. This resource played a critical role in isolating the gene responsible for some 15% of colon cancers.

Of Mice and Humans: The Value of Comparative Analyses

A remaining challenge is to recognize and discriminate all the functional constituents of a gene, particularly regulatory components not represented within cDNAs, and to predict what each gene may actually do in human biology. Comparing human and mouse sequences is an exceptionally powerful way to identify homologous genes and regulatory elements that have been substantially conserved during evolution.

Researchers led by Leroy Hood (University of Washington, Seattle) have analyzed more than 1 million bases of sequence from T-cell receptor (TCR)

To IMAGE the Human Gene Map

Since 1993, the Integrated Molecular Analysis of Gene Expression (IMAGE) Consortium has played a major role in the development of a human gene map. Founding members of the IMAGE Consortium are Bento Soares (Columbia University, now at University of Iowa), Gregory Lennon (LLNL), Mihael Polymeropoulos (National Institutes of Health's National Institute of Mental Health), and Charles Auffrey (Généthon, in France). Because cDNA molecules represent coding (expressed-gene) areas of the genome, sets of cloned cDNAs are a valuable resource to the gene-mapping community. The cDNA libraries representing different tissues have many members in common. Thus, good coordination among participating laboratories can minimize redundant work. The international IMAGE Consortium laboratories fulfill this role by developing and arraying cDNA clones for worldwide use. *[http://www-bio.llnl.gov/bbrp/image/image.html]*

From the IMAGE cDNA clones, researchers at the Washington University (St. Louis) Sequencing Center determine ESTs with support from Merck, Inc. The data, which are used in gene localization, are then entered into public databases. More than 10,000 chromosomal assignments have been entered into Genome Database (*http://www.gdb.org*). Including replica copies, over 3 million clones have been distributed, probably representing about 50,000 distinct human genes.

The IMAGE infrastructure is being used in two additional programs. At LLNL, the IMAGE laboratory arrays mouse cDNA libraries produced by Soares for the Washington University Mouse EST project (*http://genome.wustl.edu/est/mouse_esthmpg.html*) with sequencing sponsored by the Howard Hughes Medical Institute. Additional clone libraries are being used in a collaborative sequencing project sponsored by the NIH National Cancer Institute as part of the Cancer Genome Anatomy Project to identify and fully sequence genes implicated in major cancers (*http://www.ncbi.nlm.nih.gov/ncicgap*).

chromosome regions of both human and mouse genomes. Many subtle functional elements can be recognized only by comparing human and mouse sequences. TCRs play a major role in immunity and autoimmune disease, and insights into their mechanisms may one day help treat or even prevent such diseases as arthritis, diabetes, and multiple sclerosis (possibly even AIDS).

Comparative analysis is also used to model human genetic diseases. Given sequence information, researchers can produce targeted mutations in the mouse as a rapid and economical route to elucidating gene function. Such studies continue to be used effectively at Oak Ridge National Laboratory (ORNL).

DNA Sequencing

From the beginning of the genome project, DOE's DNA sequencing-technology program has supported both improvements to established methodologies and innovative higher-risk strategies. The first major sequencing project, a test bed for incremental improvements, culminated with elucidation of the highly complex TCR region (described above) by a team led by Hood.

A novel "directed" sequencing strategy initiated at LBNL in 1993 provides a potential alternative approach that can include automation as a core design feature. In this approach, every sequencing template is first mapped to its original position on a chromosome (resolution, 30 bases). The advantages of this method include a large reduction in the number of sequencing reactions needed and in the sequence-assembly steps that follow. To date, this directed strategy has achieved significant results with simpler, less repetitive nonhuman sequences, particularly in the NIH-funded *Drosophila* genome program. The system also is in use at the Stanford Human Genome Center and Mercator Genetics, Inc.

The preparation of DNA clones for sequencing involves several biochemical processing steps that require different solution environments. At the Whitehead Institute, Trevor Hawkins has improved systems for reversible binding of DNA molecules to magnetic beads that are compatible with complete robotic management. The second-generation Sequatron fits on a tabletop with a single robotic arm moving sample trays between servicing stations. This very compact system, supported by sophisticated software, may be ideal for laboratories with limited or costly floor space.

Fluorescent tags are critical components of conventional automated sequencing approaches. The team of Richard Mathies and Alexander Glazer (University of California, Berkeley) has made a series of improvements in fluorescence systems that have decreased DNA input needs and markedly increased the quality of raw data, thereby supporting longer useful reads of DNA sequence.

Complementary improvements in enzymology have been achieved by the team of Charles Richardson and Stanley Tabor (Harvard Medical School). Current widely used procedures for automated DNA sequencing involve cycling between high and low temperatures. The Harvard researchers used information about the three-dimensional structure of polymerases (enzymes needed for DNA replication) and how they function to engineer an improved Taq polymerase. ThermoSequenase, which is now produced commercially as part of the ThermoSequenase kit, reduces the amount of expensive sequencing reagents required and supports popular cycle-sequencing protocols.

The application of higher electrical fields in gel electrophoresis separation of DNA fragments can increase sequencing speed and efficiency. Conventional thick gels cannot adequately dissipate the additional heat produced, however. Two promising routes to "thinness" are ultrathin slab gels and capillary systems. An ultrathin gel system was developed by Lloyd Smith (University of Wisconsin, Madison) and licensed for commercial development.

The replacement of gels by pumpable solutions of long polymers is making capillary array electrophoresis (CAE) potentially practical for DNA sequencing. The first CAE system for DNA was demonstrated by the team of Barry Karger (Northeastern University). In 1995, Karger and Norman Dovichi (University of Alberta, Canada) separately identified CAE conditions under which DNA sequencing reads could be extended usefully up to the 1000-base range. Another CAE system, developed by Edward Yeung (Iowa State University), has been licensed for commercial production (see box, p. 23). Mathies has developed a system in which a confocal microscope displays DNA bands. Application of this system to the sizing of larger DNA fragments binding multiple fluors allows single-molecule detection.

Replacing the gel-separation step with mass spectroscopy (MS) is another promising approach for rapid DNA sequencing. MS uses differences in mass-to-charge ratios to separate ionized atoms or molecules. Early efforts at MS sequencing were plagued by chemical reactivity during the "launching" phase of matrix-assisted laser desorption ionization (MALDI). MALDI badly degraded the DNA sample input. However, the degradation chemistry was elucidated in Smith's laboratory, leading to improvements. At ORNL, the team of Chung-Hsuan Chen has performed extensive trials of alternative matrices and has achieved significant improvements that now support sequence reads up to 100 DNA bases. The system is undergoing trials for DNA diagnostic applications.

The most revolutionary sequencing technology is being pursued by the team of Richard Keller and James Jett at LANL. Their goal is to read out sequence from single DNA molecules, work that builds on LANL's expertise in flow cytometry. The strand to be sequenced is labeled first with fluors that distinguish the four DNA subunits and is then suspended in a flow stream. An exonuclease cleaves the subunits, which flow past an interrogating laser system that reports the subunits' identities. All system constituents are operational but limited by the low subunit release rates of commercially available exonucleases. A current developmental focus is on identifying more active exonucleases.

Synthetic DNA strands in the 15- to 30-base range (oligomers) play essential roles in DNA sequencing; in sample-preparation steps for the polymerase chain reaction, which copies DNA strands millions of times; and in DNA-based diagnostics. The cost of custom oligomer synthesis once was a limiting factor in many research projects. A more economical, highly parallel oligomer synthesis technology was developed by Thomas Brennan at Stanford University (see last bullet, p. 22, for further details).

The sequencing by hybridization (SBH) technology provides information only on short stretches of DNA in a single trial (interrogation), but thousands of low-cost interrogations can be performed in parallel. SBH is very useful for rapid classification of short DNAs such as cDNAs, very low cost DNA resequencing, and detection of DNA sequence differences (polymorphisms) over short regions. The team of Radomir Crkvenjakov and Radoje Drmanac invented one format of SBH while in Yugoslavia, made substantial improvements at Argonne National Laboratory (ANL), and later started Hyseq Inc. to commercialize these technologies. At ANL, another implementation, SBH on matrices (SHOM) of gels, holds promise for high-accuracy sequence proofreading and diverse DNA diagnostics. The ANL team, led by Andrei Mirzabekov, collaborates with the Englehardt Institute in Moscow, where SHOM was demonstrated initially.

Informatics: Data Collection and Analysis

Explosive growth of information and the challenges of acquiring, representing, and providing access to data pose continuing monumental tasks for the large public databases. Over the last 3 years, the Genome Database (GDB), the major international repository of human genome mapping data, has made extensive changes culminating in the enhanced representation of genomic maps and gene information in GDB V6.0. Major issues for the Genome Sequence DataBase (GSDB), established in 1994, are to capture and annotate the sequence data and to represent it in a form capable of supporting complex, ad hoc queries. Both GDB and GSDB have been restructured recently to handle the increasing flood of data and make it more useful for downstream biology (see Research Narratives, GDB, p. 49, and GSDB, p. 55. *[http://www.gdb.org* and *http://www.ncgr.org/gsdb]*

Victor Markowitz, formerly of LBNL, has developed a suite of database tools allowing substantial modifications of underlying data structures while the biologists' query tools remain stable. *[http://gizmo.lbl.gov/DM_TOOLS/DMTools.html]*

The Genome Annotation Consortium (based at ORNL) was initiated in 1997 to be a modular, distributed informatics facility for analyzing and processing (e.g., annotating) genome-scale sequence data.

The many improvements in World Wide Web software now enable maps to be downloaded simply by using a browser with accessory software provided by GDB. Computers sift stretches of DNA sequence for patterns that identify such biologically important features as protein-coding regions (exons), regulatory areas, and RNA splice sites. Other computer tools are used to compare a new sequence (i.e., a putative gene) against all other database entries, retrieve any homologous sequences that already have been entered, and indicate the degree of similarity.

The Gene Recognition and Analysis Internet Link (GRAIL) at ORNL localizes genes and other biologically important sequence features (see box, p. 17).

Another analytical service that returns informative, annotated data is MAGPIE, provided through ANL by Terry Gaasterland. MAGPIE is designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated, with automated continued reevaluation as search databases grow (*http://www.mcs.anl.gov/home/gaasterl/magpie.html*). Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact. To this end, the WIT (What is There) system, which succeeds PUMA, supports the construction of metabolic pathways. Such constructions or models are based on sequence data, the clearly established biochemistry of specific organisms, and an understanding of the interdependencies of biochemical mechanisms. WIT, which was developed by Evgenij Selkov and Ross Overbeek at ANL, offers a particularly valuable tool for testing current hypotheses about microbial biology. *[http://www.cme.msu.edu/WIT]*

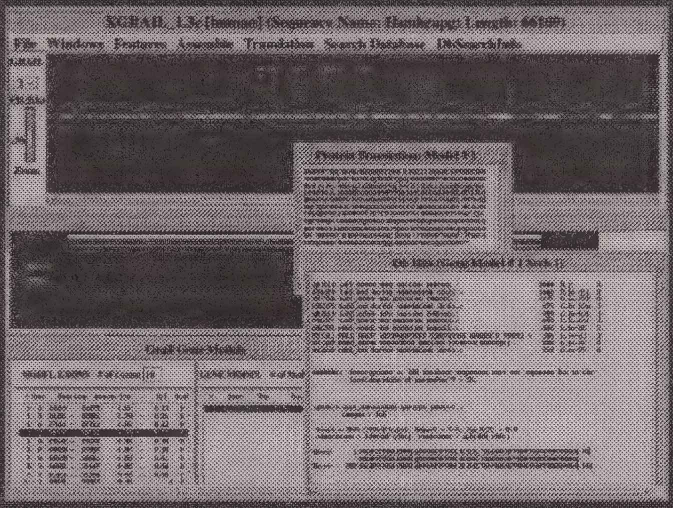
Researchers at the University of Colorado have developed another approach for predicting coding regions in genomic DNA, combining multiple types of evidence into a single scoring function, and returning both optimal and ranked suboptimal solutions. The approach is robust to substitution errors but sensitive to frameshift errors. The group is now exploring methods for predicting other classes of sequence regions, especially promoters. [software

GRAIL and GenQuest

In 1996 the Gene Recognition and Analysis Internet Link (GRAIL) processed nearly 40 million bases of sequence per month, making it the most widely used "gene-finding" system available. Developed at Oak Ridge National Laboratory (ORNL) by a team led by Ed Uberbacher, GRAIL uses artificial intelligence and machine learning to discover complex relationships in sequence data. The genQuest server, also at ORNL, compares information generated by GRAIL with data in protein, DNA, and motif databases to add further value to annotation of DNA sequences.

GRAIL's latest version (1.3) combines a Motif Graphical Client with improved sensitivity and splice-site recognition, better performance in AT-rich regions, new analysis systems for model organisms, and frameshift detection. This system can be used on a wide variety of UNIX platforms, including Sun, DEC, and SGI. The many ways to access GRAIL include a command line sockets client that permits remote program calls to all basic GRAIL-genQuest analysis services, thus allowing convenient integration of GRAIL results into automated analysis pipelines.

Contact GRAIL staff through the Web site at *http://compbio.ornl.gov* or at *GRAILMAIL@ornl.gov* for e-mail and ftp access.



The figure above shows the GRAIL analysis of part of the human major histocompatibility locus, which carries genes responsible for cellular immunity. Included in this analysis are potential exons (gene-coding regions), gene models, CpG islands (areas rich in bases C and G found in most mammalian genes), and repetitive DNA elements. (Source: Richard Mural, ORNL)

and information: *http://beagle.colorado.edu/~eesnyder/GeneParser.html]*

The Baylor College of Medicine (BCM) Search Launcher improves user access to the wide variety of database-search tools available on the Web. Search Launcher features a single point of entry for related searches, the addition of hypertext links to results returned by remote servers, and a batch client. *[http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html]*

FASTA-SWAP, also from the BCM group, is a new pattern-search tool for databases that improves sensitivity and specificity to help detect related sequences. BEAUTY, an enhanced version of the BLAST database-search program, improves access to information about the functions of matched sequences and incorporates additional hypertext links. Graphical displays allow correlation of hit positions with annotated domain positions. Future plans include providing access to information from and direct links to other databases, including organism-specific databases.

PROCRUSTES uses comparisons of the same gene of different species to delimit gene structure much more accurately. The product of a collaboration between Pavel Pevzner (University of Southern California) and two Russian researchers, PROCRUSTES is based on the spliced-alignment algorithm, which explores all possible exon assemblies and finds the multiexon structure that best fits a related protein. *[http://www-hto.usc.edu/software/procrustes]*

The Ethical, Legal, and Social Issues component of the DOE Human Genome Program supports projects to help judges understand the scientific validity of the genetics-based claims that are poised to flood the nation's courtrooms. Robert F. Orr (left) of the North Carolina Supreme Court and Francis X. Spina of the Massachusetts Appeals Court at the New England Regional Conference on the Courts and Genetics (July 1997) participate in a hands-on laboratory session. As a prelude to learning the fundamentals of DNA science and genetic testing, the judges are precipitating DNA (seen as streaks on the glass rod in the tube) from a solution containing the bacterium Escherichia coli. *[Courts and Science On-Line Magazine:* http://www.ornl.gov/courts*]*

Ethical, Legal, and Social Issues (ELSI)

From the outset of the Human Genome Project, researchers recognized that the resulting increase in knowledge about human biology and personal genetic information would raise complex ethical and policy issues for individuals and society. Rapid worldwide progress in the project has heightened the urgency of this challenge.

Most observers agree that personal knowledge of genetic susceptibility can be expected to serve humankind well, opening the door to more accurate diagnoses, preventive intervention, intensified screening, lifestyle changes, and early and effective treatment. But such knowledge has another side, too: risk of anxiety, unwelcome changes in personal relationships, and the danger of stigmatization. Often, genetic tests can indicate possible future medical conditions far in advance of any symptoms or available therapies or treatments. If handled carelessly, genetic information could threaten an individual with discrimination by potential employers and insurers.

Other issues are perhaps less immediate than these personal concerns but no less challenging. How, for example, are products of the Human Genome Project to be patented and commercialized? How are the judicial, medical, and educational communities—not to mention the public at large—to be educated effectively about genetic research and its implications?

To confront these issues, the DOE and NIH ELSI programs jointly established an ELSI working group to coordinate policy and research between the two agencies. *[*An FY 1997 report evaluating the joint ELSI group is available on the Web (*http://www.ornl.gov/hgmis/archive/elsirept.html*).*]*

The DOE Human Genome Program has focused its ELSI efforts on education, privacy, and the fair use of genetic information (including ownership and commercialization); workplace issues, especially screening for susceptibilities to environmental agents; and implications of research findings regarding interactions among multiple genes and environmental influences.

A few highlights from the DOE ELSI portfolio for FY 1994 through FY 1997 are outlined below.

- Three high school curriculum modules developed by the Biological Sciences Curriculum Study (BSCS). *[http://www.bscs.org]*
- An educational program in Los Angeles to develop a culturally and linguistically appropriate genetics curriculum based on a BSCS module (see above) for Hispanic students and their families. *[http://vflylab.calstatela.edu/hgp]*
- A series of workshops to educate a core group of 1000 judges around the nation and a handbook with companion videotape to assist federal and state judges in understanding and assessing genetic evidence in an increasing number of civil and criminal cases (see photo above).

- Educational materials developed by the Science+Literacy for Health Project of the American Association for the Advancement of Science (AAAS) and targeted at or above the 6th- to 8th-grade reading levels. [AAAS: 202/326-6453; *Your Genes, Your Choices* booklet: *http://www.nextwave.org/ehr/books/index.html]*
- A program at the University of Chicago aimed at developing a knowledge base for physicians and nurses who will train other practitioners to introduce new genetic services.
- A series of radio programs (see photo at right) on the science and ethical issues of the genome project and a TV documentary program on ELSI issues. *[http://www.pbs.org]*
- *The Gene Letter*, a monthly online newsletter on ELSI issues for healthcare professionals and consumers. *[http://www.geneletter.org]*
- A congressional fellowship program in human genetics, administered through AAAS, for one annual fellowship for a mid-career geneticist. *[society@genetics.faseb.org]*
- The draft Genetic Privacy Act, prepared as a model for privacy legislation and covering the collection, analysis, storage, and use of DNA samples and the genetic information derived from them. *[http://www.ornl.gov/hgmis/resource/privacy/privacy1.html]*
- Privacy studies at the Center for Social and Legal Research, including an analysis of the effects of new genetic technologies on individuals and institutions.

For details on these and other projects, see ELSI Abstracts, p. 45, in Part 2 of this report. In addition to the specific projects listed in Part 2, the DOE program sponsors a number of conferences and workshops on ELSI topics.

THE DNA FILES

Leroy Hood (left) of the University of Washington, Seattle, talks with Bart Scott at the 1996 DOE Human Genome Program Contractor-Grantee Workshop. Scott represented the Genome Radio Project (see text at left), which is supported by the Ethical, Legal, and Social Issues Program of the DOE Human Genome Program. (See the project's abstract in Part 2 of this report for more information.)

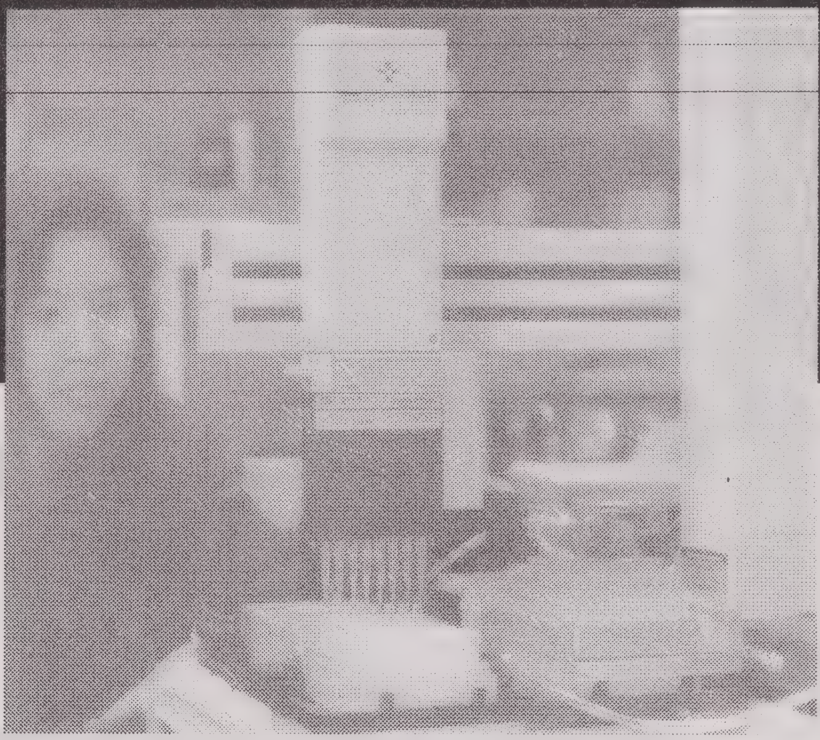
DOE ELSI Web Site
http://www.ornl.gov/hgmis/resource/elsi.html

Protection of Human Research Subjects

In 1996, President Clinton appointed the National Bioethics Advisory Commission to provide guidance on the ethical conduct of current and future biological and behavioral research, especially that related to genetics and the rights and welfare of human research subjects (*http://www.nih.gov/nbac/nbac.htm*).

Also in 1996, DOE and NIH issued a document providing investigators with guidance in the use of DNA from human subjects for large-scale sequencing projects (see Appendix C: Human Subjects Guidelines, p. 77). *[http://www.ornl.gov/hgmis/archive/nchgrdoe.html]*

Lawrence Livermore National Laboratory researcher Maria de Jesus, who designed software to automate DNA isolation. (Source: Linda Ashworth, LLNL)



Technology Transfer

Converting scientific knowledge into commercially useful products

Transferring technology to the private sector, a primary mission of DOE, is strongly encouraged in the Human Genome Program to enhance the nation's investment in research and technological competitiveness. Human genome centers at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL) provide opportunities for private companies to collaborate on joint projects or use laboratory resources. These opportunities include access to information (including databases), personnel, and special facilities; informal research collaborations; Cooperative Research and Development Agreements (CRADAs); and patent and software licensing. For information on recently developed resources, contact individual genome research centers or see Research Highlights, beginning on p. 9. Many universities have their own licensing and technology transfer offices.

Some collaborations and technology-transfer highlights from FY 1994 through FY 1996 are described below.

Collaborations

Involvement of the private sector in research and development can facilitate successful transfer of technology to the marketplace, and collaborations can speed production of essential tools for genome research. A number of interactive projects are now under way, and others are in preliminary stages.

CRADAs

One technology-transfer mechanism used by DOE national laboratories is the CRADA, a legal agreement with a nongovernmental organization to collaborate on a defined research project. Under a CRADA, the two entities share scientific and technological expertise, with the governmental organization providing personnel, services, facilities, equipment, or other resources. Funds must come from the nongovernmental partner. A benefit to participating companies is the opportunity to negotiate exclusive licenses for inventions arising from these collaborations. For periods through 1996, the CRADAs in place in the DOE Human Genome Program included the following:

- LLNL with Applied Biosystems Division of Perkin-Elmer Corporation to develop analytical instrumentation for faster DNA sequencing instrumentation;
- LANL with Amgen, Inc., to develop bioassays for cell growth factors;
- Oak Ridge National Laboratory (ORNL) with Darwin Molecular, Inc., for mouse models of human immunologic disease;
- ORNL with Proctor & Gamble, Inc., for analyses of liver regeneration in a mouse model; and
- Brookhaven National Laboratory with U.S. Biochemical Corporation to identify proteins useful for primer-walking methods and large-scale sequencing.

Work for Others

In other collaborations, the LBNL genome center is participating in a Work for Others agreement with Amgen to automate the isolation and characterization of large numbers of mouse cDNAs. The center group is focusing on adapting LBNL's automated colony-picking system to cDNA protocols and applying methods to generate large numbers of filter replicas for colony

Technology Transfer Legislation

Technology transfer involves converting scientific knowledge into commercially useful products. Through the 1980s, a series of laws was enacted to encourage the development of commercial applications of federally funded research at universities and federal laboratories. Such laws [chiefly the Bayh-Dole Act of 1980, Stevenson-Wydler Act of 1980, and Federal Technology Transfer Act of 1986 (Public Laws 96-517, 96-480, and 99-502, respectively)] were not aimed specifically at genome or even biomedical research. However, such research and the surrounding commercial biotechnology enterprises clearly have benefited from them. The biotechnology sector's success owes much to federal policies on technology transfer and intellectual property. [Source: U.S. Congress, Office of Technology Assessment, *Federal Technology Transfer and the Human Genome Project*, OTA-BP-EHR-162 (Washington, D.C.: U.S. Government Printing Office, September 1995)] filter hybridization and subsequent analysis. ["Work for Others" projects supported by an agency or organization other than DOE (e.g., NIH, National Cancer Institute, or a private company) can be conducted at a DOE installation because this work is complementary to DOE research missions and usually requires multidisciplinary DOE facilities and technologies.]

The Resource for Molecular Cytogenetics was established at LBNL and the University of California (UC), San Francisco, with the support of the Office of Biological and Environmental Research and Vysis, Inc. (formerly Imagenetics). The Resource aims to apply fluorescent in situ hybridization (FISH) techniques to genetic analysis of human tissue samples; produce probe reagents; design and develop digital-imaging microscopy; distribute probes, analysis technology, and educational materials in the molecular cytogenetic community; and transfer useful reagents, processes, and instruments to the private sector for commercialization.

NIST Advanced Technology Program

Several commercial applications of research sponsored by the U.S. Human Genome Project have been furthered by the Advanced Technology Program (ATP) of the U.S. National Institute of Standards and Technology. ATP's mission is to stimulate economic growth and industrial competitiveness by encouraging high-risk but powerful new technologies. Its Tools for DNA Diagnostics program uses collaborations among researchers and industry to develop (1) cost-effective methods for determining, analyzing, and storing DNA sequences for a wide variety of diagnostic applications ranging from healthcare to agriculture to the environment and (2) a new and potentially very large market for DNA diagnostic systems.

Awardees have included companies developing DNA diagnostic chips, more powerful cytogenetic diagnostic techniques based on comparative genomic hybridization, DNA sequencing instrumentation, and DNA analysis technology. Eventually, commercialization of these underlying technologies is expected to generate hundreds of thousands of jobs. *[800/287-3863, Fax: 301/926-9524, atp@micf.nist.gov, http://www.atp.nist.gov]*

Patenting and Licensing Highlights, FY 1994–96

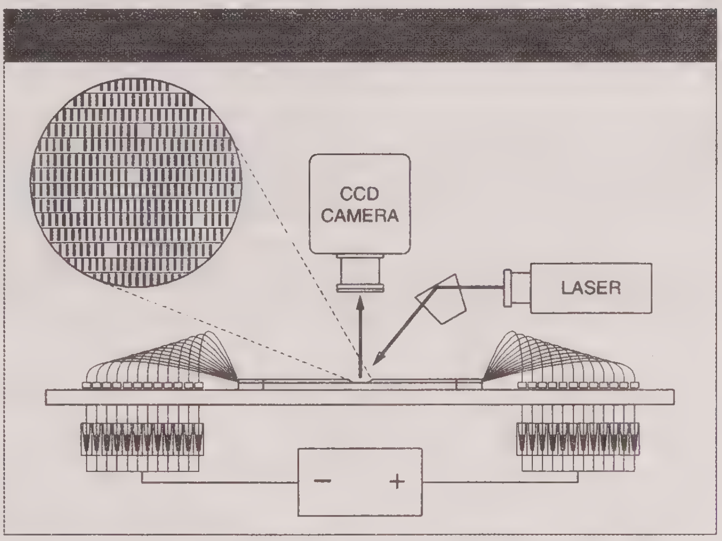
- A development license for single-molecule DNA sequencing replaced the 1991–94 CRADA (the first CRADA to be established in the U.S. Human Genome Project) between LANL and Life Technologies, Inc. (LTI).
- In 1995, a broad patent was awarded to UC for chromosome painting. This technology uses FISH to stain specific locations in cells and chromosomes for diagnosing, imaging, and studying chromosomal abnormalities and cancer. Resulting from a 1989 CRADA between LLNL and UC, FISH was licensed exclusively to Vysis.
- Hyseq, Inc., was founded in 1993 by former Argonne National Laboratory researchers Radoje Drmanac and Radomir Crkvenjakov to commercialize the sequencing by hybridization (SBH) technology. Hyseq has exclusive patent rights to a variation known as format 3 of SBH or the "super chip." Hyseq later won an Advanced Technology Program award from the U.S. National Institute of Standards and Technology to develop the technology further.
- Oligomers—short, single-stranded DNAs—are crucial reagents for genome research and biomedical diagnostics. ProtoGene Laboratories, Inc., was founded to commercialize new DNA synthesis technology (developed initially at LBNL with completed prototypes at Stanford University) and to offer the first lower-cost custom oligomer synthesis. The Parallel Array Synthesis system, which independently synthesizes 96 oligomers per run in a standard 96-well microtiter plate format, shows great promise for significant cost reductions. ProtoGene first

licensed sales and distribution to LTI and, later, production rights as well. LTI operates production centers in the United States, Europe, and Japan.

- The GRAIL-genQuest sequence-analysis software developed at ORNL was licensed by Martin Marietta Energy Systems (now Lockheed Martin Energy Research) to ApoCom, Inc., for pharmaceutical and biotechnology company researchers who cannot use the Internet because of data-security concerns. The public GRAIL-genQuest service remains freely available on the Internet (see box, p. 17).
- In 1995, an exclusive license was granted to U.S. Biochemical Corporation for a genetically engineered, heat-stable, DNA-replicating enzyme with much-improved sequencing properties. The enzyme was developed by Stanley Tabor at Harvard University Medical School.
- In 1995, an advanced capillary array electrophoresis system for sequencing DNA was patented by Iowa State University. The system was licensed to Premier American Technologies Corporation for commercialization (see graphic at right and R&D 100 Awards, next page).
- In 1996, a patent was granted to LANL researchers for DNA fragment sizing and sorting by laser-induced fluorescence. An exclusive license was awarded to Molecular Technology, Inc., for commercialization of the single-molecule detection capability related to DNA sizing (see R&D 100 Awards, next page).

SBIR and STTR

Small Business Innovation Research (SBIR) Program awards are designed to stimulate commercialization of new technology for the benefit of both the private and public sectors. The highly competitive program emphasizes cutting-edge, high-risk research with potential for high payoff in different areas, including human genome research. Small business firms with fewer than 500 employees are invited to submit applications. SBIR human genome topics concentrate on innovative and experimental approaches for carrying out the goals of the Human Genome Project (see SBIR, p. 63, in Part 2 of this report). The Small Business Technology Transfer (STTR) Program fosters transfers between research institutions and small businesses. [DOE SBIR and ST.TR contact: Kay Etzler (301/903-5867, Fax: -5488, *Kay.Etzler@oer.doe.gov*), *http://sbir.er.doe.gov/sbir, http://sttr.er.doe.gov/sttr]*



Capillary Array Electrophoresis (CAE). *CAE systems promise dramatically faster and higher-resolution fragment separation for DNA sequencing. A multiplexed CAE system designed by Edward Yeung (Iowa State University) has been developed for commercial production by Premier American Technologies Corporation (PATCO). In the PATCO ESY9600 model, DNA samples are introduced into the 96-capillary array; as the separated fragments pass through the capillaries, they are irradiated all at once with laser light. Fluorescence is measured by a charged coupled device that acts as a simultaneous multichannel detector. (Inset circle at upper left: Closeup view of individual capillary lanes with separated samples.) Because every fragment length exists in the sample, bases are identified in order according to the time required for them to reach the laser-detector region.*

[Source: Thomas Kane, PATCO]

Technology Transfer Award

A Federal Laboratory Consortium Award for Excellence in Technology Transfer was presented to Edward Yeung and a research team at Iowa State University's Ames Laboratory in 1993. Their laser-based method for indirect fluorescence of biological samples may have applications for routine high-speed DNA sequencing (see graphic, p. 23). Yeung also won the 1994 American Chemical Society Award for Analytical Chemistry.

1997 R&D 100 Awards

DOE researchers in 12 facilities across the country won 36 of the R&D 100 Awards given by *Research and Development Magazine* for 1996 work. DOE award-winning research ranged from advances in supercomputing to the biological recycling of tires. Announced in July 1997, these awards bring DOE's R&D 100 total to 453, the most of any single organization and twice as many as all other government agencies combined.

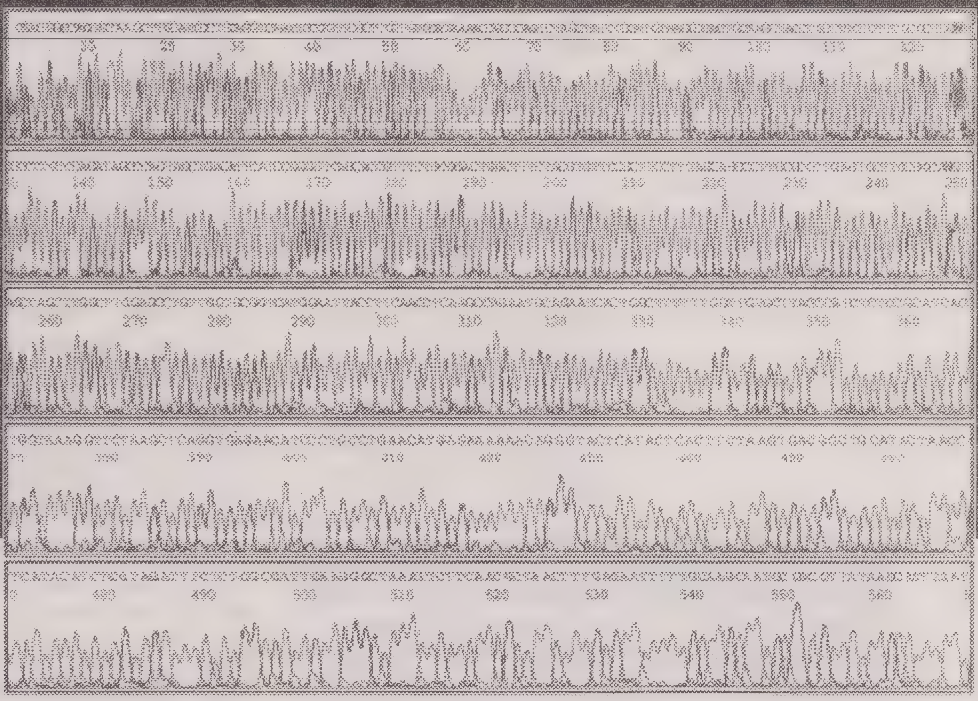
Two DOE genome-related research projects received 1997 R&D 100 Awards. One was to Yeung (see text at left and graphic, p. 23) for "ESY9600 Multiplexed Capillary Electrophoresis DNA Sequencer."

The other award was to Richard Keller and James Jett (LANL) with Amy Gardner (Molecular Technologies, Inc.) for "Rapid-Size Analysis of Individual DNA Fragments." This technology speeds determination of DNA fragment sizes, making DNA fingerprinting applications in biotechnology and other fields more reliable and practical.

R&D Magazine began making annual awards in 1963 to recognize the 100 most significant new technologies, products, processes, and materials developed throughout the world during the previous year (*http://www.rdmag.com/rd100/100award.htm*). Winners are chosen by the magazine's editors and a panel of 75 respected scientific experts in a variety of disciplines. Previous winners of R&D 100 Awards include such well-known products as the flashcube (1965), antilock brakes (1969), automated teller machine (1973), fax machine (1975), digital compact cassette (1993), and Taxol anticancer drug (1993).

Research Narratives

Readout from an automated DNA sequencing machine depicts the order of the four DNA bases (A, T, C, and G) in a DNA fragment of more than 500 bases. (Source: Linda Ashworth, LLNL)



Joint Genome Institute 26
Lawrence Livermore National Laboratory 27
Los Alamos National Laboratory 35
Lawrence Berkeley National Laboratory 41
University of Washington Genome Center 47
Genome Database 49
National Center for Genome Resources 55

Joint Genome Institute

DOE Merges Sequencing Efforts of Genome Centers

http://www.jgi.doe.gov

Elbert Branscomb
JGI Scientific Director
Lawrence Livermore
National Laboratory
7000 East Avenue, L-452
Livermore, CA 94551
510/422-5681
elbert@alu.llnl.gov or
elbert@shotgun.llnl.gov

In a major restructuring of its Human Genome Program, on October 23, 1996, the DOE Office of Biological and Environmental Research established the Joint Genome Institute (JGI) to integrate work based at its three major human genome centers.

The JGI merger represents a shift toward large-scale sequencing via intensified collaborations for more effective use of the unique expertise and resources at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (see Research Narratives, beginning on p. 27 in this report). Elbert Branscomb (LLNL) serves as JGI's Scientific Director. Capital equipment has been ordered, and operational support of about $30 million is projected for the 1998 fiscal year.

With easy access to both LBNL and LLNL, a building in Walnut Creek, California, is being modified. Here, starting in late FY 1998, production DNA sequencing will be carried out for JGI. Until that time, large-scale sequencing will continue at LANL, LBNL, and LLNL. Expectations are that within 3 to 4 years the Production Sequencing Facility will house some 200 researchers and technicians working on high-throughput DNA sequencing using state-of-the-art robotics.

Initial plans are to target gene-rich regions of around 1 to 10 megabases for sequencing. Considerations include gene density, gene families (especially clustered families), correlations to model organism results, technical capabilities, and relevance to the DOE mission (e.g., DNA repair, cancer susceptibility, and impact of genotoxins). The JGI program is subject to regular peer review.

Sequence data will be posted daily on the Web; as the information progresses to finished quality, it will be submitted to public databases.

As JGI and other investigators involved in the Human Genome Project are beginning to reveal the DNA sequence of the 3 billion base pairs in a reference human genome, the data already are becoming valuable reagents for explorations of DNA sequence function in the body, sometimes called "functional genomics." Although large-scale sequencing is JGI's major focus, another important goal will be to enrich the sequence data with information about its biological function. One measure of JGI's progress will be its success at working with other DOE laboratories, genome centers, and non-DOE academic and industrial collaborators. In this way, JGI's evolving capabilities can both serve and benefit from the widest array of partners.

Production DNA Sequencing Begun Worldwide

The year 1996 marked a transition to the final and most challenging phase of the U.S. Human Genome Project, as pilot programs aimed at refining large-scale sequencing strategies and resources were funded by DOE and NIH (see Research Highlights, DNA Sequencing, p. 14). Internationally, large-scale human genome sequencing was kicked off in late 1995 when The Wellcome Trust announced a 7-year, $75-million grant to the private Sanger Centre to scale up its sequencing capabilities. French investigators also have announced intentions to begin production sequencing.

Funding agencies worldwide agree that rapid and free release of data is critical. Other issues include sequence accuracy, types of annotation that will be most useful to biologists, and how to sustain the reference sequence.

The international Human Genome Organisation maintains a Web page to provide information on current and future sequencing projects and links to sites of participating groups (*http://hugo.gdb.org*). The site also links to reports and resources developed at the February 1996 and 1997 Bermuda meetings on large-scale human genome sequencing, which were sponsored by The Wellcome Trust.

Lawrence Livermore National Laboratory Human Genome Center

http://www-bio.llnl.gov/bbrp/genome/genome.html

Human Genome Center
Lawrence Livermore National
Laboratory
Biology and Biotechnology
Research Program
7000 East Avenue, L-452
Livermore, CA 94551

Anthony V. Carrano
Director
510/422-5698, Fax: /423-3110
carrano1@llnl.gov

Linda Ashworth
Assistant to Center Director
510/422-5665, Fax: -2282
ashworth1@llnl.gov

The Human Genome Center at Lawrence Livermore National Laboratory (LLNL) was established by DOE in 1991. The center operates as a multidisciplinary team whose broad goal is understanding human genetic material. It brings together chemists, biologists, molecular biologists, physicists, mathematicians, computer scientists, and engineers in an interactive research environment focused on mapping, DNA sequencing, and characterizing the human genome.

Goals and Priorities

In the past 2 years, the center's goals have undergone an exciting evolution. This change is the result of several factors, both intrinsic and extrinsic to the Human Genome Project. They include: (1) successful completion of the center's first-phase goal, namely a high-resolution, sequence-ready map of human chromosome 19; (2) advances in DNA sequencing that allow accelerated scaleup of this operation; and (3) development of a strategic plan for LLNL's Biology and Biotechnology Research Program that will integrate the center's resources and strengths in genomics with programs in structural biology, individual susceptibility, medical biotechnology, and microbial biotechnology.

The primary goal of LLNL's Human Genome Center is to characterize the mammalian genome at optimal resolution and to provide information and material resources to other in-house or collaborative projects that allow exploitation of genomic biology in a synergistic manner. DNA sequence information provides the biological driver for the center's priorities:

- Generation of highly accurate sequence for chromosome 19.
- Generation of highly accurate sequence for genomic regions of high biological interest to the mission of the DOE Office of Biological and Environmental Research (e.g., genes involved in DNA repair, replication, recombination, xenobiotic metabolism, and cell-cycle control).
- Isolation and sequence of the full insert of cDNA clones associated with genomic regions being sequenced.
- Sequence of selected corresponding regions of the mouse genome in parallel with the human.
- Annotation and position of the sequenced clones with physical landmarks such as linkage markers and sequence tagged sites (STSs).
- Generation of mapped chromosome 19 and other genomic clones [cosmids, bacterial artificial chromosomes (BACs), and P1 artificial chromosomes (PACs)] for collaborating groups.
- Sharing of technology with other groups to minimize duplication of effort.
- Support of downstream biology projects, for example, structural biology, functional studies, human variation, transgenics, medical biotechnology, and microbial biotechnology with know-how, technology, and material resources.

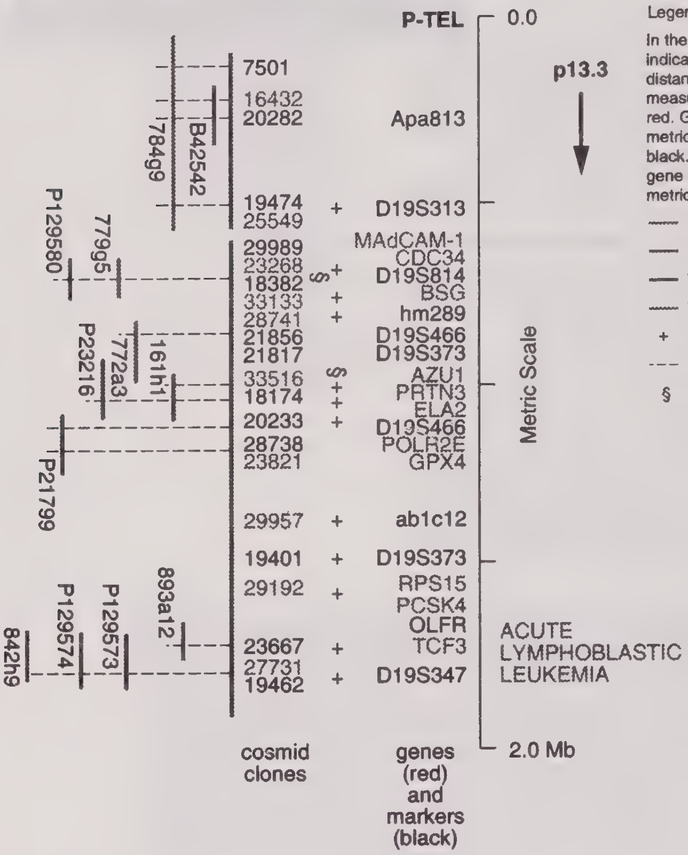
Center Organization and Activities

Completion and publication of the metric physical map of human chromosome 19 (see p. 28) in 1995 has led to consolidation of many functions associated with physical mapping, with increased emphasis on DNA sequencing. The center is organized into five broad areas of research and support: sequencing, resources, functional genomics, informatics and analytical genomics, and instrumentation. Each area consists of multiple projects, and extensive interaction occurs both within and among projects.

In lieu of individual abstracts, research projects and investigators at the LLNL Human Genome Center are represented in this narrative. More information can be found on the center's Web site (see URL above).

Update

In 1997 Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory began collaborating in a Joint Genome Institute to implement high-throughput sequencing [see p. 26 and *Human Genome News* 8(2), 1–2].


P-TEL
0.0
p13.3
7501
16432
20282
Apa813
784g9
B42542
19474
25549
D19S313
29989
23268
18382
33133
28741
21856
21817
33516
18174
20233
28738
23821
MAdCAM-1
CDC34
D19S814
BSG
hm289
D19S466
D19S373
AZU1
PRTN3
ELA2
D19S466
POLR2E
GPX4
P129580
779g5
P23216
772a3
161h1
P21799
Metric Scale
29957
ab1c12
19401
D19S373
29192
RPS15
PCSK4
OLFR
893a12
P129573
P129574
842h9
23667
TCF3
27731
19462
D19S347
ACUTE LYMPHOBLASTIC LEUKEMIA
2.0 Mb
cosmid clones
genes (red) and markers (black)

Legend

In the column labeled cosmid clones, black indicates a FISH-ordered clone where distance between clones has been measured. Other cosmids are shown in red. Genes are in red to the left of the metric scale. Other markers are labeled in black. A disease associated with a specific gene is shown in blue to the right of the metric scale.

- —— Restriction-mapped contig
- —— BAC, PAC, or P1 clone
- —— YAC with known and concordant size
- —— YAC with unknown or discordant size
- + Sequence tagged site (STS)
- --- STS and/or hybridization results
- § Polymorphic marker

Chromosome 19 Map. *In the current map (at left) of the first 2 million bases at the p-telomere end of chromosome 19, the* EcoR I *restriction-mapped contigs (represented by red lines) provide the starting material for genomic sequencing across a region.*

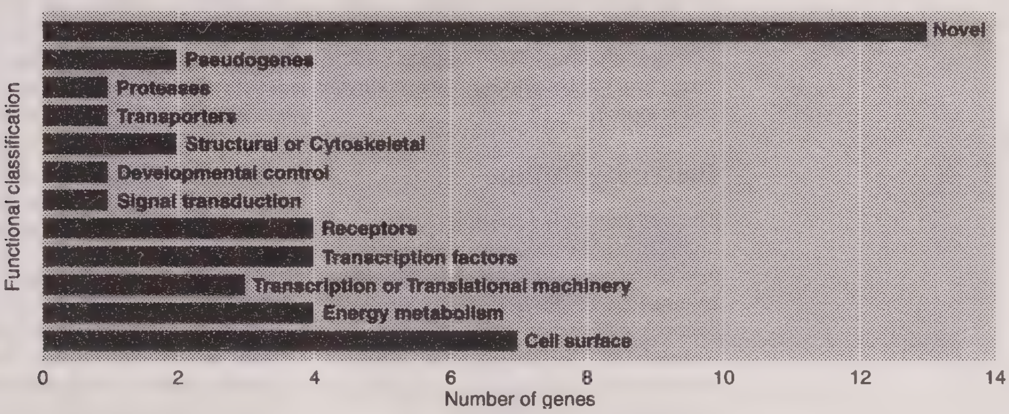
Construction of the human chromosome 19 physical map was based on a similar strategy for mapping the roundworm Caenorhabditis elegans. *View the complete map on the World Wide Web* (http://www-bio.llnl.gov/genome/html/chrom_map.html). *[Source: Adapted from figure provided by Linda Ashworth, LLNL]*

Sequencing

The sequencing group is divided into several subprojects. The core team is responsible for the construction of sequence libraries, sequencing reactions, and data collection for all templates in the random phase of sequencing. The finishing team works with data produced by the core team to produce highly redundant, highly accurate "finish" sequence on targets of interest. Finally, a team of researchers focuses specifically on development, testing, and implementation of new protocols for the entire group, with an emphasis on improving the efficiency and cost basis of the sequencing operation.

Resources

The resources group provides mapped clonal resources to the sequencing teams. This group performs physical mapping as needed for the DNA sequencing group by using fingerprinting, restriction mapping, fluorescence in situ hybridization, and other techniques. A small mapping effort is under way to identify, isolate, and characterize BAC



Putative-Gene Classification. *The figure depicts the functional classification of putative genes identified in a 1.02-Mb region on the long arm of human chromosome 19. Analysis of the completed sequence between markers D19S208 and COX7A1 revealed 43 open reading frames (ORFs) or putative genes. (An ORF is a DNA region containing specific sequences that signal the beginning and ending of a gene.)*

Thirty of these putative genes were found to have sequence similarities to a wide variety of known genes or proteins, including some involved in transcription, cell adhesion and signaling, and metabolism. Many appear to be related functionally to such known proteins as the GTP-ase activating proteins or the ETS family of transcription factors. Others seem to be new members of existing gene families, for example, the mRNA splicing factor. or of such pseudogenes as the elongation factor Tu.

In addition to those that could be classified, 13 novel genes were identified, including one with high similarity to a predicted ORF of unknown function in the roundworm Caenorhabditis elegans. *[Source: Adapted from graph provided by Linda Ashworth, LLNL]*

clones (from anywhere in the human genome) that relate to susceptibility genes, for example, DNA repair. These clones will be characterized and provided for sequencing and at the same time contribute to understanding the biology of the chromosome, the genome, and susceptibility factors. The mapping team also collaborates with others using the chromosome 19 map as a resource for gene hunting.

Functional Genomics

The functional genomics team is responsible for assembling and characterizing clones for the Integrated Molecular Analysis of Gene Expression (called IMAGE) Consortium and cDNA sequencing, as well as for work on gene expression and comparative mouse genomics. The effort emphasizes genes involved in DNA repair and links strongly to LLNL's gene-expression and structural biology efforts. In addition, this team is working closely with Oak Ridge National Laboratory (ORNL) to develop a comparative map and the sequence data for mouse regions syntenic to human chromosome 19 (see p. 32).

Informatics and Analytical Genomics

The informatics and analytical genomics group provides computer science support to biologists. The sequencing informatics team works directly with the DNA sequencing group to facilitate and automate sample handing, data acquisition and storage, and DNA sequence analysis and annotation. The analytical genomics team provides statistical and advanced algorithmic expertise. Tasks include development of model-based methods for data capture, signal processing, and feature extraction for DNA sequence and fingerprinting data and analysis of the effectiveness of newly proposed methods for sequencing and mapping.

Instrumentation

The instrumentation group also has multiple components. Group members provide expertise in instrumentation and automation in high-throughput electrophoresis, preparation of high-density replicate DNA and colony filters, fluorescence labeling technologies, and automated sample handling for DNA sequencing. To facilitate seamless integration of new technologies into production use, this group is coupled tightly to the biologist user groups and the informatics group.

Collaborations

The center interacts extensively with other efforts within the LLNL Biology and Biotechnology Research Program and with other programs at LLNL, the academic community, other research institutes, and industry. More than 250 collaborations range from simple probe and clone sharing to detailed gene family studies. The following list reflects some major collaborations.

- Integration of the genetic map of human chromosome 19 with corresponding mouse chromosomes (ORNL).
- Miniaturized polymerase chain reaction instrumentation (LLNL).
- Sequencing of IMAGE Consortium cDNA clones (Washington University, St. Louis).
- Mapping and sequencing of a gene associated with Finnish congenital nephrotic syndrome (University of Oulu, Finland).

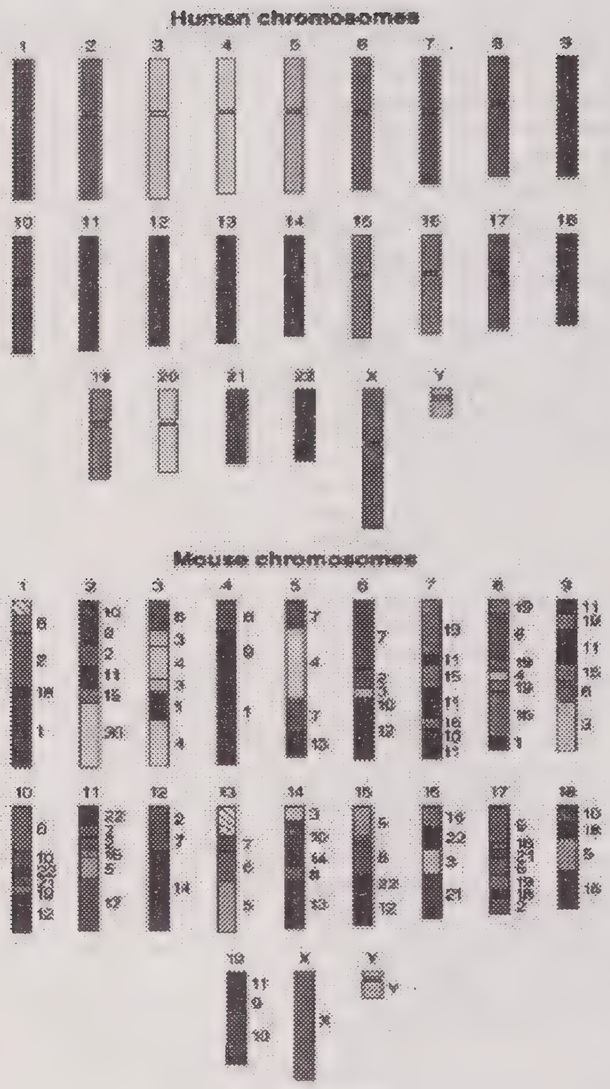
Accomplishments

The LLNL Human Genome Center has excelled in several areas, including comparative genomic sequencing of DNA repair genes in human and rodent species, construction of a metric physical map of human chromosome 19, and development and application of new biochemical and mathematical approaches for constructing ordered clone maps. These and other major accomplishments are highlighted below.

- Completion of highly accurate sequencing totaling 1.6 million bases of DNA, including regions spanning human DNA repair genes, the candidate region for a congenital kidney disease gene, and other regions of biological interest on chromosome 19.
- Completion of comparative sequence analysis of 107,500 bases of genomic DNA encompassing the human DNA repair gene *ERCC*2 and the corresponding regions in mouse and hamster (p. 32). In addition to *ERCC*2, analysis revealed the presence of two previously undescribed genes in all three species. One of these genes is a new member of the kinesin motor protein family. These proteins play a wide variety of roles in the cell, including movement of chromosomes before cell division.
- Complete sequencing of human genomic regions containing two additional DNA repair genes. One of these, *XRCC*3, maps to human chromosome 14 and encodes a protein that may be required for chromosome stability. Analysis of the genomic sequence identified another kinesin motor protein gene physically linked to *XRCC*3. The second human repair gene, *HHR*23A, maps to 19p13.2. Sequence analysis of 110,000 bases containing *HHR*23A identified six other genes, five of which are new genes with similarity

to proteins from mouse, human, yeast, and *Caenorhabditis elegans*.

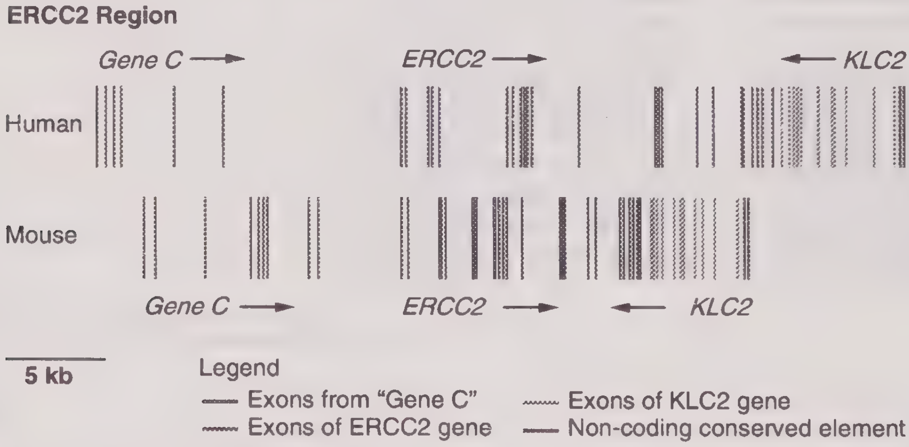
- Complete sequencing of full-length cDNAs for three new DNA repair genes (*XRCC*2, *XRCC*3, and *XRCC*9) in collaboration with the LLNL DNA repair group.
- Generation of a metric physical map of chromosome 19 spanning at least 95% of the chromosome. This unique map incorporates a metric scale to estimate the distance between genes or other markers of interest to the genetics community.
- Assembly of nearly 45 million bases of *Eco*R I restriction-mapped cosmid contigs for human chromosome 19 using a combination of fingerprinting and cosmid walking. Small gaps in cosmid continuity have been spanned by BAC, PAC, and P1 clones, which are then integrated into the restriction maps. The high depth of coverage of these maps (average redundancy, 4.3-fold) permits selection of a minimum overlapping set of clones for DNA sequencing.
- Placement of more than 400 genes, genetic markers, and other loci on the chromosome 19 cosmid map. Also, 165 new STSs associated with premapped cosmid contigs were generated and added to the physical map.
- Collaborations to identify the gene (*COMP*) responsible for two allelic genetic diseases, pseudoachondroplasia and multiple epiphyseal dysplasia, and the identification of specific mutations causing each condition.
- Through sequence analysis of the 2A subfamily of the human cytochrome P450 enzymes, identification of a new variant that exists in 10% to 20% of individuals and results in reduced ability to metabolize nicotine and the antiblood-clotting drug Coumadin.
- Location of a zinc finger gene that encodes a transcription factor regulating blood-cell development adjacent to telomere repeat sequences, possibly the gene nearest one end of chromosome 19.
- Completion of the genomic and cDNA sequence of the gene for the human Rieske Fe-S protein involved in mitochondrial respiration.
- Expansion of the mouse-human comparative genomics collaboration with ORNL to include study of new groups of clustered transcription factors found on human chromosome 19q and as syntenic homologs on mouse chromosome 7 (p. 32).
- Numerous collaborations (in particular, with Washington University and Merck) continuing to expand the LLNL-based IMAGE Consortium, an effort to characterize the transcribed human genome. The IMAGE clone collection is now the largest public collection of sequenced cDNA clones, with more than one million arrayed clones, 800,000 sequences in public databases, and 10,000 mapped cDNAs.
- Development and deployment of a comprehensive system to handle sample tracking needs of production DNA sequencing. The system combines databases and graphical interfaces running on both Mac and Sun platforms and scales easily to handle large-scale production sequencing.
- Expansion of the LLNL genome center's World Wide Web site to include tables that link to each gene being sequenced, to the quality scores and assembled bases collected each night during the sequencing process, and to the submitted GenBank sequence when a clone is completed. *[http://bbrp.llnl.gov/test-bin/projqcsummary]*


Human chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Mouse chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Y

Human-Mouse Homologies. *LLNL researcher Lisa Stubbs (above) is shown in the Mouse Genetics Research Facility at ORNL. [ORNL photo]*

The figure at left demonstrates the genetic similarity (homology) of the superficially dissimilar mouse and human species. The similarity is such that human chromosomes can be cut (schematically at least) into about 150 pieces (only about 100 are large enough to appear here), then reassembled into a reasonable approximation of the mouse genome. The colors and corresponding numbers on the mouse chromosomes indicate the human chromosomes containing homologous segments. [Source: Lisa Stubbs, LLNL]

Comparative sequencing of homologous regions in human and mouse at LLNL has enhanced the ability to identify protein-coding (exon) and noncoding DNA regions that have remained unchanged over the course of evolution. Colors in the figure below depict similarities in mouse and human genes involved in DNA repair, a research interest rooted in DOE's mission to develop better technologies for measuring health effects, particularly mutations. [Source: Linda Ashworth, LLNL]


ERCC2 Region
Gene C →
ERCC2 →
← KLC2
Human
Mouse
Gene C →
ERCC2 →
← KLC2
5 kb
Legend
Exons from "Gene C"
Exons of ERCC2 gene
Exons of KLC2 gene
Non-coding conserved element

- Implementation of a new database to support sequencing and mapping work on multiple chromosomes and species. Web-based automated tools were developed to facilitate construction of this database, the loading of over 100 million bytes of chromosome 19 data from the existing LLNL database, and automated generation of Web-based input interfaces.
- Significant enhancement of the LLNL Genome Graphical Database Browser software to display and link information obtained at a subcosmid resolution from both restriction map hybridization and sequence feature data. Features, such as genes linked to diseases, allow tracking to fragments as small as 500 base pairs of DNA.
- Development of advanced microfabrication technologies to produce electrophoresis microchannels in large glass substrates for use in DNA sequencing.
- Installation of a new filter-spotting robot that routinely produces 6 × 6 × 384 filters. A 16 × 16 × 384 pattern has been achieved.
- Upgrade of the Lawrence Berkeley National Laboratory colony picker using a second computer so that imaging and picking can occur simultaneously.

Future Plans

Genomic sequencing currently is the dominant function of Livermore's Human Genome Center. The physical mapping effort will ensure an ample supply of sequence-ready clones. For sequencing targets on chromosome 19, this includes ensuring that the most stable clones (cosmids, BACs, and PACs) are available for sequencing and that regions with such known physical landmarks as STSs and expressed sequenced tags (ESTs) are annotated to facilitate sequence assembly and analysis. The following targets are emphasized for DNA sequencing:

- Regions of high gene density, including regions containing gene families.
- Chromosome 19, of which at least 42 million bases are sequence ready.
- Selected BAC and PAC clones representing regions of about 0.2 million to 1 million bases throughout the human genome; clones would be selected based on such high-priority biological targets as genes involved in DNA repair, replication, recombination, xenobiotic metabolism, cell-cycle checkpoints, or other specific targets of interest.
- Selected BAC and PAC clones from mouse regions syntenic with the genes indicated above.
- Full-insert cDNAs corresponding to the genomic DNA being sequenced.

The informatics team is continuing to deploy broader-based supporting databases for both mapping and sequencing. Where appropriate, Web- and Java-based tools are being developed to enable biologists to interact with data. Recent reorganization within this group enables better direct support to the sequencing group, including evaluating and interfacing sequence-assembly algorithms and analysis tools, data and process tracking, and other informatics functions that will streamline the sequencing process.

The instrumentation effort has three major thrusts: (1) continued development or implementation of laboratory automation to support high-throughput sequencing; (2) development of the next-generation DNA sequencer; and (3) development of robotics to support high-density BAC clone screening. The last two goals warrant further explanation.

The new DNA sequencer being developed under a grant from the National Institutes of Health, with minor support through the DOE genome center, is designed to run 384 lanes simultaneously with a low-viscosity sieving medium. The entire system would be loaded automatically, run, and set up for the next run at 3-hour intervals. If successful, it should provide a 20- to 40-fold increase in throughput over existing machines.

An LLNL-designed high-precision spotting robot, which should allow a density of 98,304 spots in 96 cm^2, is now operating. The goal of this effort is to create high-density filters representing a 10× BAC coverage of both human and mouse genomes (30,000 clones = 1× coverage). Thus each filter would provide ~3× coverage, and eight such filters would provide the desired coverage for both genomes. The filters would be hybridized with amplicons from individual or region-specific cDNAs and ESTs; given the density of the BAC libraries, clones that hybridize should represent a binned set of BACs for a region of interest. These BACs could be the initial substrate for a BAC sequencing strategy. Performing hybridizations in parallel in mouse and human DNA facilitates the development of the mouse map (with ORNL involvement), and sequencing BACs from both species identifies evolutionarily conserved and, perhaps, regulatory regions.

Information generated by sequencing human and mouse DNA in parallel is expected to expand LLNL efforts in functional genomics. Comparative sequence data will be used to develop a high-resolution synteny map of conserved mouse-human domains and incorporate automated northern expression analysis of newly identified genes. Long range, the center hopes to take advantage of a variety of forms of expression analysis, including site-directed mutation analysis in the mouse.

Summary

The Livermore Human Genome Center has undergone a dramatic shift in emphasis toward commitment to large-scale, high-accuracy sequencing of chromosome 19, other chromosomes, and targeted genomic regions in the human and mouse. The center also is committed to exploiting sequence information for functional genomics studies and for other programs, both in house and collaboratively.

Los Alamos National Laboratory Center for Human Genome Studies

http://www-ls.lanl.gov/masterhgp.html

Center for Human Genome Studies
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545

Larry L. Deaven
Acting Director
505/667-3912, Fax: -2891
ldeaven@telomere.lanl.gov

Lynn Clark
Technical Coordinator
505/667-9376, Fax: -2891
clark@telomere.lanl.gov

Robert K. Moyzis
Director, 1989–97*

Biological research was initiated at Los Alamos National Laboratory (LANL) in the 1940s, when the laboratory began to investigate the physiological and genetic consequences of radiation exposure. Eventual establishment of the national genetic sequence databank called GenBank, the National Flow Cytometry Resource, numerous related individual research projects, and fulfillment of a key role in the National Laboratory Gene Library Project all contributed to LANL's selection as the site for the Center for Human Genome Studies in 1988.

Center Organization and Activities

The LANL genome center is organized into four broad areas of research and support: Physical Mapping, DNA Sequencing, Technology Development, and Biological Interfaces. Each area consists of a variety of projects, and work is distributed among five LANL Divisions (Life Sciences; Theoretical; Computing, Information, and Communications; Chemical Science and Technology; and Engineering Sciences and Applications). Extensive interdisciplinary interactions are encouraged.

Physical Mapping

The construction of chromosome- and region-specific cosmid, bacterial artificial chromosome (BAC), and yeast artificial chromosome (YAC) recombinant DNA libraries is a primary focus of physical mapping activities at LANL. Specific work includes the construction of high-resolution maps of human chromosomes 5 and 16 and associated informatics and gene discovery tasks.

Accomplishments

- Completion of an integrated physical map of human chromosome 16 consisting of both a low-resolution YAC contig map and a high-resolution cosmid contig map (pp. 37–39). With sequence tagged site (STS) markers provided on average every 125,000 bases, the YAC-STS map provides almost-complete coverage of the chromosome's euchromatic arms. All available loci continue to be incorporated into the map.
- Construction of a low-resolution STS map of human chromosome 5 consisting of 517 STS markers regionally assigned by somatic-cell hybrid approaches. Around 95% mega-YAC–STS coverage (50 million bases) of 5p has been achieved. Additionally, about 40 million bases of 5q mega-YAC–STS coverage have been obtained collaboratively.
- Refinement of BAC cloning procedures for future production of chromosome-specific libraries. Successful partial digestion and cloning of microgram quantities of chromosomal DNA embedded in agarose plugs. Efforts continue to increase the average insert size to about 100,000 bases.

DNA Sequencing

DNA sequencing at the LANL center focuses on low-pass sample sequencing (SASE) of large genomic regions. SASE data is deposited in publicly available databases to allow for wide distribution. Finished sequencing is prioritized from initial SASE analysis and pursued by parallel primer walking. Informatics development includes data tracking, gene-discovery integration with the Sequence Comparison ANalysis (SCAN) program, and functional genomics interaction.

Accomplishments

- SASE sequencing of 1.5 million bases from the p13 region of human chromosome 16.
- Discovery of more than 100 genes in SASE sequences.

In lieu of individual abstracts, research projects and investigators at the LANL Center for Human Genome Studies are represented in this narrative. More information can be found on the center's Web site (see URL above).

Update

In 1997 Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory began collaborating in a Joint Genome Institute to implement high-throughput sequencing [see p. 26 and *Human Genome News* 8(2), 1–2].

---

*Now at University of California, Irvine

- Generation of finished sequence for a 240,000-base telomeric region of human chromosome 7q. From initial sequences generated by SASE, oligonucleotides were synthesized and used for primer walking directly from cosmids comprising the contig map. Complete sequencing was performed to determine what genes, if any, are near the 7q terminus. This intriguing region lacks significant blocks of subtelomeric repeat DNA typically present near eukaryotic telomeres.
- Complete single-pass sequencing of 2018 exon clones generated from LANL's flow-sorted human chromosome 16 cosmid library. About 950 discrete sequences were identified by sequence analysis. Nearly 800 appear to represent expressed sequences from chromosome 16.
- Development of Sequence Viewer to display ABI sequences with trace data on any computer having an Internet connection and a Netscape World Wide Web browser.
- Sequencing and analysis of a novel pericentromeric duplication of a gene-rich cluster between 16p11.1 and Xq28 (in collaboration with Baylor College of Medicine).

Technology Development

Technology development encompasses a variety of activities, both short and long term, including novel vectors for library construction and physical mapping; automation and robotics tools for physical mapping and sequencing; novel approaches to DNA sequencing involving single-molecule detection; and novel approaches to informatics tools for gene identification.

Accomplishments

- Development of SCAN program for large-scale sequence analysis and annotation, including a translator converting SCAN data to GIO format for submission to Genome Sequence DataBase.
- Application of flow-cytometric approach to DNA sizing of P1 artificial chromosome (PAC) clones. Less than one picogram of linear or supercoiled DNA is analyzed in under 3 minutes. Sizing range has been extended down to 287 base pairs. Efforts continue to extend the upper limit beyond 167,000 bases.
- Characterization of the detection of single, fluorescently tagged nucleotides cleaved from multiple DNA fragments suspended in the flow stream of a flow cytometer (see picture, p. 70). The cleavage rate for Exo III at 37°C was measured to be about 5 base pairs per second per M13 DNA fragment. To achieve a single-color sequencing demonstration, either the background burst rate (currently about 5 bursts per second) must be reduced or the exonuclease cleavage rate must be increased significantly. Techniques to achieve both are being explored.
- Construction of a simple and compact apparatus, based on a diode-pumped Nd:YAG laser, for routine DNA fragment sizing.
- Development of a new approach to detect coding sequences in DNA. This complete spectral analysis of coding and noncoding sequences is as sensitive in its first implementations as the best existing techniques.
- Use of phylogenetic relationships to generate new profiles of amino acid usage in conserved domains. The profiles are particularly useful for classification of distantly related sequences.

Biological Interfaces

The Biological Interfaces effort targets genes and chromosome regions associated with DNA damage and repair, mitotic stability, and chromosome structure and function as primary subjects for physical mapping and sequencing. Specific disease-associated genes on human chromosome 5 (e.g., Cri-du-Chat syndrome) and on 16 (e.g., Batten's disease and Fanconi anemia) are the subjects of collaborative biological projects.

Accomplishments

- Identification of two human 7q exons having 99% homology to the cDNA of a known human gene, vasoactive intestinal peptide receptor 2A. Preliminary data suggests that the *VIPR*2A gene is expressed.
- Identification of numerous expressed sequence tags (ESTs) localized to the 7q region. Since three of the ESTs contain at least two regions with high confidence of homology (~90%), genes in addition to *VIPR*2A may exist in the terminal region of 7q.
- Generation of high-resolution cosmid coverage on human chromosome 5p for the larynx and critical regions identified with Cri-du-Chat syndrome, the most common human terminal-deletion syndrome (in collaboration with Thomas Jefferson University).
- Refinement of the Wolf-Hirschhorn syndrome (WHS) critical region on human chromosome 4p. Using the SCAN program to identify genes likely to contribute to WHS, the project serves as a model for defining the interaction between genomic sequencing and clinical research.
- Collaborative construction of contigs for human chromosome 16, including 1.05 million bases in cosmids through the familial Mediterranean fever (FMF) gene region (with members of the FMF Consortium) and 700,000 bases in P1 clones encompassing the polycystic kidney disease gene (with Integrated Genetics, Inc.).
- Collaborative identification and determination of the complete genomic structure of the Batten's disease gene (with members of the BDG Consortium), the gamma subunit of the human amiloride-sensitive epithelial channel (Liddle's syndrome, with University of Iowa), and the polycystic kidney disease gene (with Integrated Genetics).
- Participation in an international collaborative research consortium that successfully identified the gene responsible for Fanconi anemia type A.

Chromosome 16 Physical Map (pp. 38–39). *A condensed chromosome 16 physical map constructed at Los Alamos National Laboratory (LANL) is shown in two parts on the following pages. Besides facilitating the isolation and characterization of disease genes, the map provides the framework for a large-scale sequencing effort by LANL, The Institute for Genomic Research, and the Sanger Centre.*

Distinct types of maps and data are shown as levels or tiers on the integrated map. At the top of each page is a view of the banded human chromosome to which the map is aligned. A somatic-cell hybrid breakpoint map, which divides the chromosome into 90 intervals, was used as a backbone for much of the map integration.

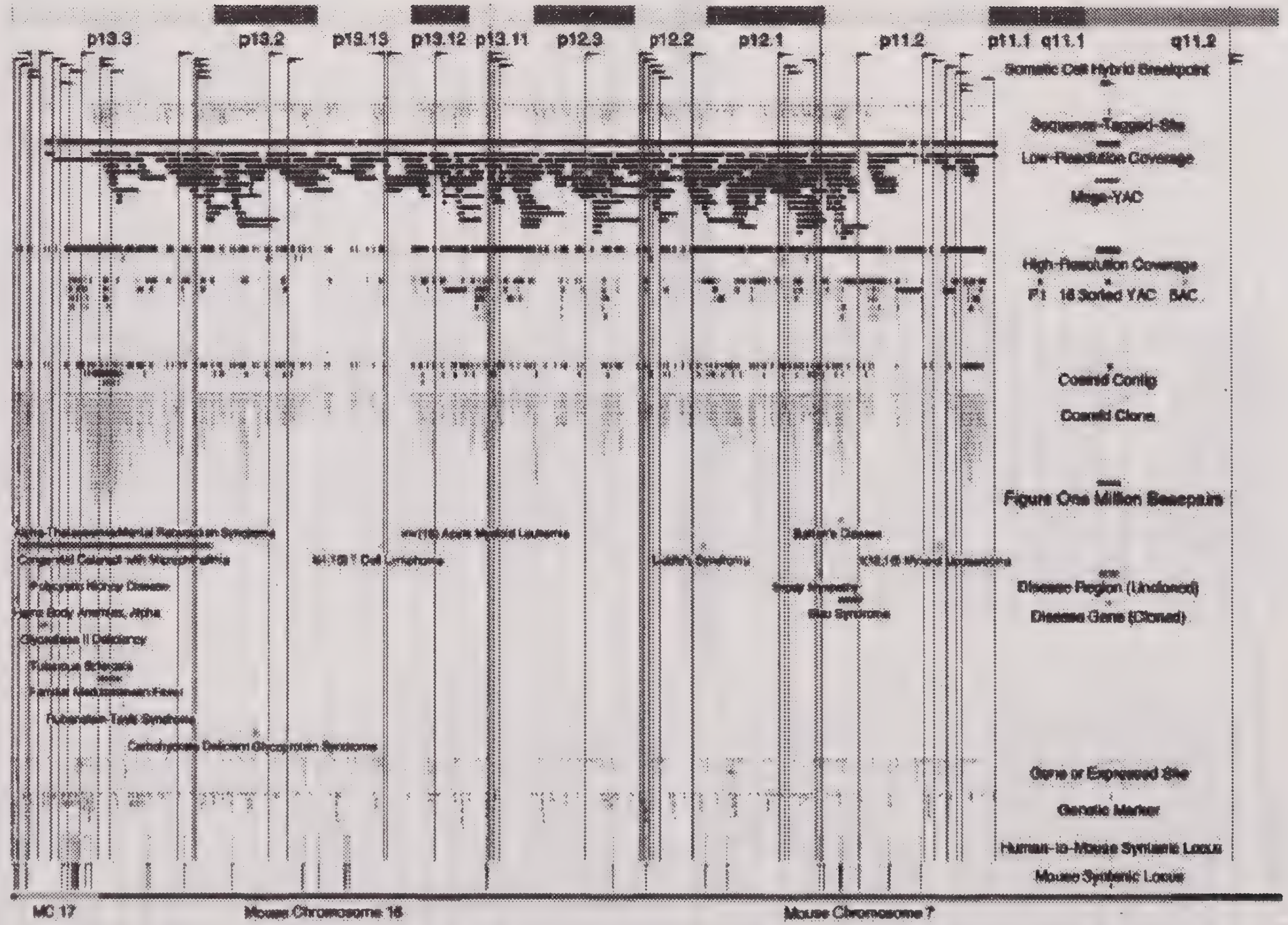
The physical map consists of both a low-resolution yeast artificial chromosome (YAC) contig map localized to and ordered within the breakpoint intervals with sequence tagged sites (STSs) and a high-resolution bacteria-based clone map. The YAC-STS map provides almost complete coverage of the chromosome's euchromatic arm, with STS markers on average every 100,000 bases.

A high-resolution, sequence-ready cosmid contig map is anchored to the YAC and breakpoint maps via STSs developed from cosmid contigs and by hybridizations between YACs and cosmids.

As part of the ongoing effort to incorporate all available loci onto a single map of this chromosome, the integrated map also features genes, expressed sequence tags, exons (gene-coding regions), and genetic markers.

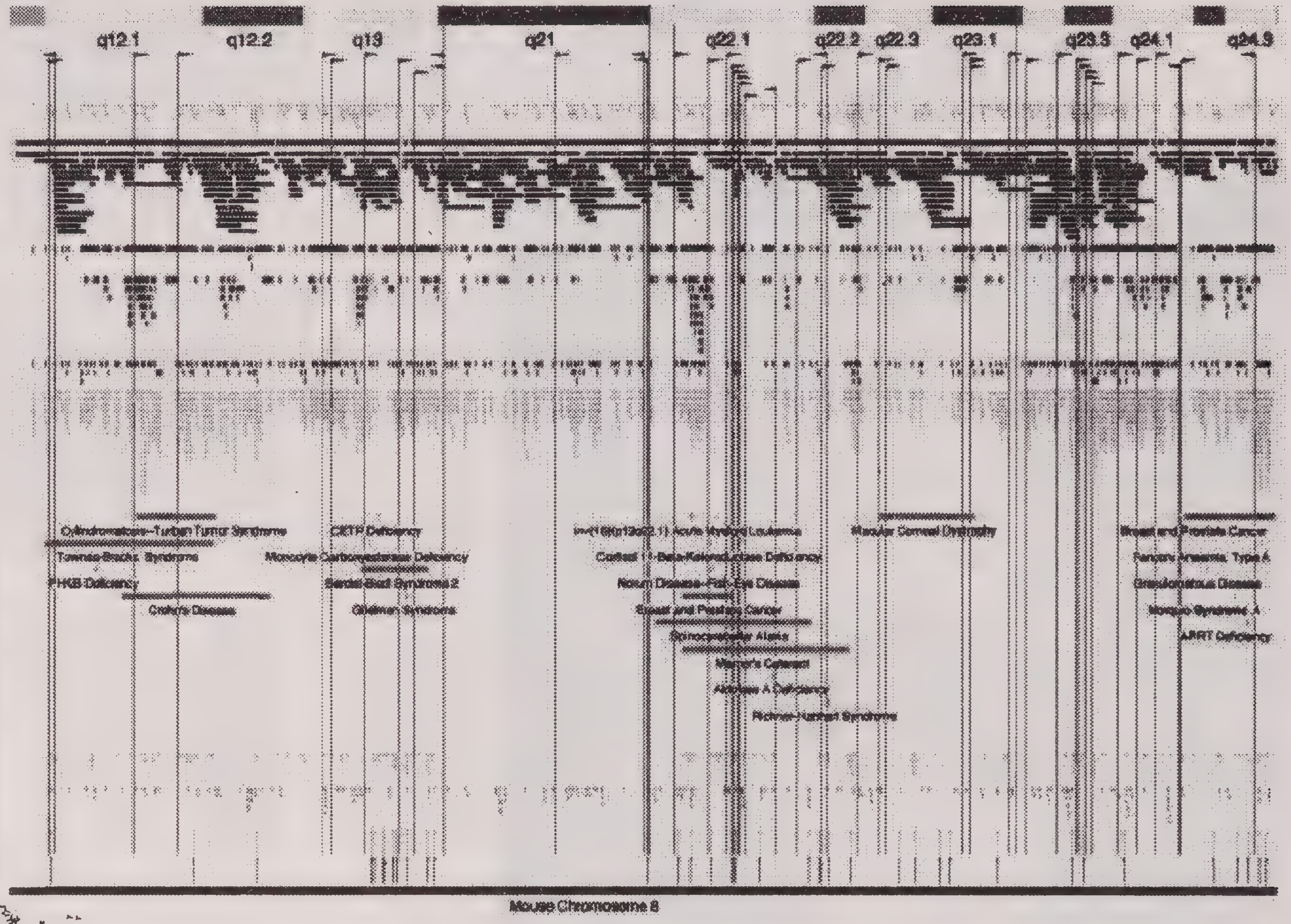
The mouse chromosome segments at the bottom of the map contain groups that correspond to human genes mapped to the regions shown above them. *[Source: Norman Doggett, LANL]*

Human Chromosome 16


p13.3
p13.2
p13.13
p13.12
p13.11
p12.3
p12.2
p12.1
p11.2
p11.1
q11.1
q11.2
Somatic Cell Hybrid Breakpoint
Sequence-Tagged-Site
Low-Resolution Coverage
Mega-YAC
High-Resolution Coverage
F1
16 Sorted YAC
BAC
Cosmid Contig
Cosmid Clone
Figure One Million Basepairs
Disease Region (Uncloned)
Disease Gene (Cloned)
Gene or Expressed Site
Genetic Marker
Human-to-Mouse Syntenic Locus
Mouse Syntenic Locus
Alpha-Thalassemia/Mental Retardation Syndrome
Congenital Cataract with Microphthalmia
Polycystic Kidney Disease
Heinz Body Anemias, Alpha
Tuberous Sclerosis
Familial Mediterranean Fever
Rubinstein-Taybi Syndrome
Carbohydrate Deficient Glycoprotein Syndrome
inv(16) Acute Myeloid Leukemia
T Cell Lymphoma
Batten's Disease
Blau Syndrome
Myxoid Liposarcoma
MC 17
Mouse Chromosome 16
Mouse Chromosome 7

38 DOE Human Genome Program Report, LANL

Human Chromosome 16


q12.1
q12.2
q13
q21
q22.1
q22.2
q22.3
q23.1
q23.3
q24.1
q24.3
Cylindromatosis–Turban Tumor Syndrome
Townes-Brocks Syndrome
PHKB Deficiency
Crohn's Disease
CETP Deficiency
Monocyte Carboxyesterase Deficiency
Bardet-Biedl Syndrome 2
Gitelman Syndrome
inv(16)(p13q22.1) Acute Myeloid Leukemia
Cortisol 11-Beta-Ketoreductase Deficiency
Norum Disease–Fish-Eye Disease
Breast and Prostate Cancer
Spinocerebellar Ataxia
Marner's Cataract
Aldolase A Deficiency
Richner-Hanhart Syndrome
Macular Corneal Dystrophy
Breast and Prostate Cancer
Fanconi Anemia, Type A
Granulomatous Disease
Morquio Syndrome, A
APRT Deficiency
Mouse Chromosome 8

HUMAN GENOME PROJECT

The exhibit "Understanding Our Genetic Inheritance" at the Bradbury Science Museum in Los Alamos, New Mexico, describes the LANL Center for Human Genome Studies' contributions to the Human Genome Project. The exhibit's centerpiece is a 16-foot-long version of LANL's map of human chromosome 16. [Source: LANL Center for Human Genome Studies]

Patents, Licenses, and CRADAs

- Rhett L. Affleck, James N. Demas, Peter M. Goodwin, Jay A. Schecker, Ming Wu, and Richard A. Keller, "Reduction of Diffusional Defocusing in Hydrodynamically Focused Flows by Complexing with a High Molecular Weight Adduct," United States Patent, filed December 1996.
- R.L. Affleck, W.P. Ambrose, J.D. Demas, P.M. Goodwin, M.E. Johnson, R.A. Keller, J.T. Petty, J.A. Schecker, and M. Wu, "Photobleaching to Reduce or Eliminate Luminescent Impurities for Ultrasensitive Luminescence Analysis," United States Patent, S-87, 208, accepted September 1997.
- J.H. Jett, M.L. Hammond, R.A. Keller, B.L. Marrone, and J.C. Martin, "DNA Fragment Sizing and Sorting by Laser-Induced Fluorescence," United States Patent, S.N. 75,001, allowed May 1996.
- James H. Jett, "Method for Rapid Base Sequencing in DNA and RNA with Three Base Labeling," in preparation.
- Development license and exclusive license to LANL's DNA sizing patent obtained by Molecular Technology, Inc., for commercialization of single-molecule detection capability to DNA sizing.

Future Plans

LANL has joined a collaboration with California Institute of Technology and The Institute for Genomic Research to construct a BAC map of the *p* arm of human chromosome 16 and to complete the sequence of a 20-million–base region of this map.

In its evolving role as part of the new DOE Joint Genome Institute, LANL will continue scaleup activities focused on high-throughput DNA sequencing. Initial targets include genes and DNA regions associated with chromosome structure and function, syntenic breakpoints, and relevant disease-gene loci.

A joint DNA sequencing center was established recently by LANL at the University of New Mexico. This facility is responsible for determining the DNA sequence of clones constructed at LANL, then returning the data to LANL for analysis and archiving.

Lawrence Berkeley National Laboratory Human Genome Center

http://www-hgc.lbl.gov/GenomeHome.html

Human Genome Center
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720

Contact:
Mohandas Narla
510/486-7029, Fax: -6746
mohandas_narla@macmail.lbl.gov

Joyce Pfeiffer
Administrative Assistant

Michael Palazzolo*
Director, 1996–97

Since 1937 the Ernest Orlando Lawrence Berkeley National Laboratory (LBNL) has been a major contributor to knowledge about human health effects resulting from energy production and use. That was the year John Lawrence went to Berkeley to use his brother Ernest's cyclotrons to launch the application of radioactive isotopes in biological and medical research. Fifty years later, Berkeley Lab's Human Genome Center was established.

Now, after another decade, an expansion of biological research relevant to Human Genome Project goals is being carried out within the Life Sciences Division, with support from the Information and Computing Sciences and Engineering divisions. Individuals in these research projects are making important new contributions to the key fields of molecular, cellular, and structural biology; physical chemistry; data management; and scientific instrumentation. Additionally, industry involvement in this growing venture is stimulated by Berkeley Lab's location in the San Francisco Bay area, home to the largest congregation of biotechnology research facilities in the world.

In July 1997 the Berkeley genome center became part of the Joint Genome Institute (see p. 26).

Sequencing

Large-scale genomic sequencing has been a central, ongoing activity at Berkeley Lab since 1991. It has been funded jointly by DOE (for human genome production sequencing and technology development) and the NIH National Human Genome Research Institute [for sequencing the *Drosophila melanogaster* model system, which is carried out in partnership with the University of California, Berkeley (UCB)]. The human genome sequencing area at Berkeley Lab consists of five groups: Bioinstrumentation, Automation, Informatics, Biology, and Development. Complementing these activities is a group in Life Sciences Division devoted to functional genomics, including the transgenics program.

The directed DNA sequencing strategy at Berkeley Lab was designed and implemented to increase the efficiency of genomic sequencing (see figure, p. 45). A key element of the directed approach is maintaining information about the relative positions of potential sequencing templates throughout the entire sequencing process. Thus, intelligent choices can be made about which templates to sequence, and the number of selected templates can be kept to a minimum. More important, knowledge of the interrelationship of sequencing runs guides the assembly process, making it more resistant to difficulties imposed by repeated sequences. As of July 3, 1997, Berkeley Lab had generated 4.4 megabases of human sequence and, in collaboration with UCB, had tallied 7.6 megabases of *Drosophila* sequence.

Instrumentation and Automation

The instrumentation and automation program encompasses the design and fabrication of custom apparatus to facilitate experiments, the programming of laboratory robots to automate repetitive procedures, and the development of (1) improved hardware to extend the applicability range of existing commercial robots and (2) an integrated operating system to control and monitor experiments. Although some discrete instrumentation modules used in the integrated protocols are obtained commercially, LBNL designs its own custom instruments when existing capabilities are inadequate. The instrumentation modules are then integrated into a large system to facilitate large-scale production sequencing. In addition, a significant effort is devoted to improving

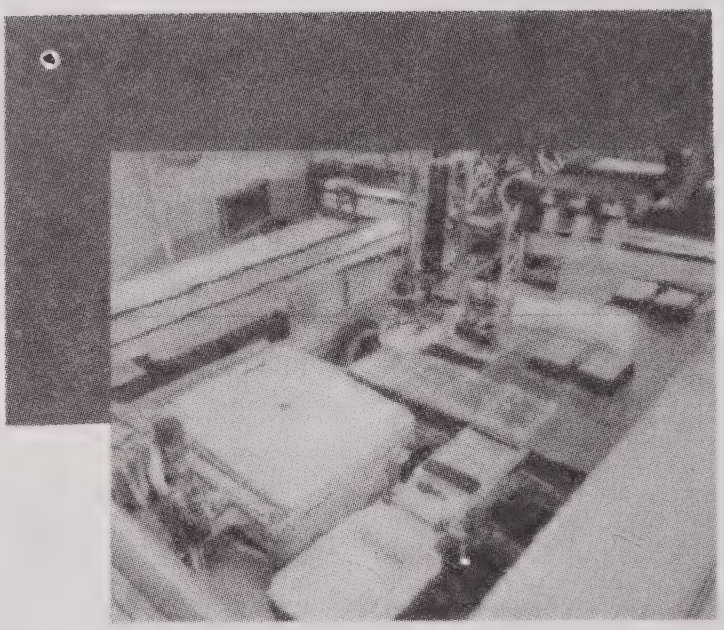
In lieu of individual abstracts, research projects and investigators at the LBNL Human Genome Center are represented in this narrative. More information can be found on the center's Web site (see URL above).

Update

In 1997 Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory began collaborating in a Joint Genome Institute to implement high-throughput sequencing [see p. 26 and *Human Genome News* 8(2), 1–2].

---

*Now at Amgen, Inc.



DNA Prep Machine. The DNA Prep machine (above) was designed by Berkeley Lab's Martin Pollard to perform plasmid preparation on 192 samples (2 microtiter plates) in about 2.5 to 4 hours, depending on the protocol. Controlled by a personal computer running a Visual Basic Control program, the instrument includes a gantry robot equipped with pipettors, reagent dispensers, hot and cold temperature stations, and a pneumatic gripper. (Source: LBNL)

fluorescence-assay methods, including DNA sequence analysis and mass spectrometry for molecular sizing.

Recent advances in the instrumentation group include DNA Prep machine and Prep Track. These instruments are designed to automate completely the highly repetitive and labor-intensive DNA-preparation procedure to provide higher daily throughput and DNA of consistent quality for sequencing (see photos, p. 43, and Web pages: *http://hgighub.lbl.gov/esd/DNAPrep/TitlePage.html* and *http://hgighub.lbl.gov/esd/prepTrackWebpage/preptrack.htm*).

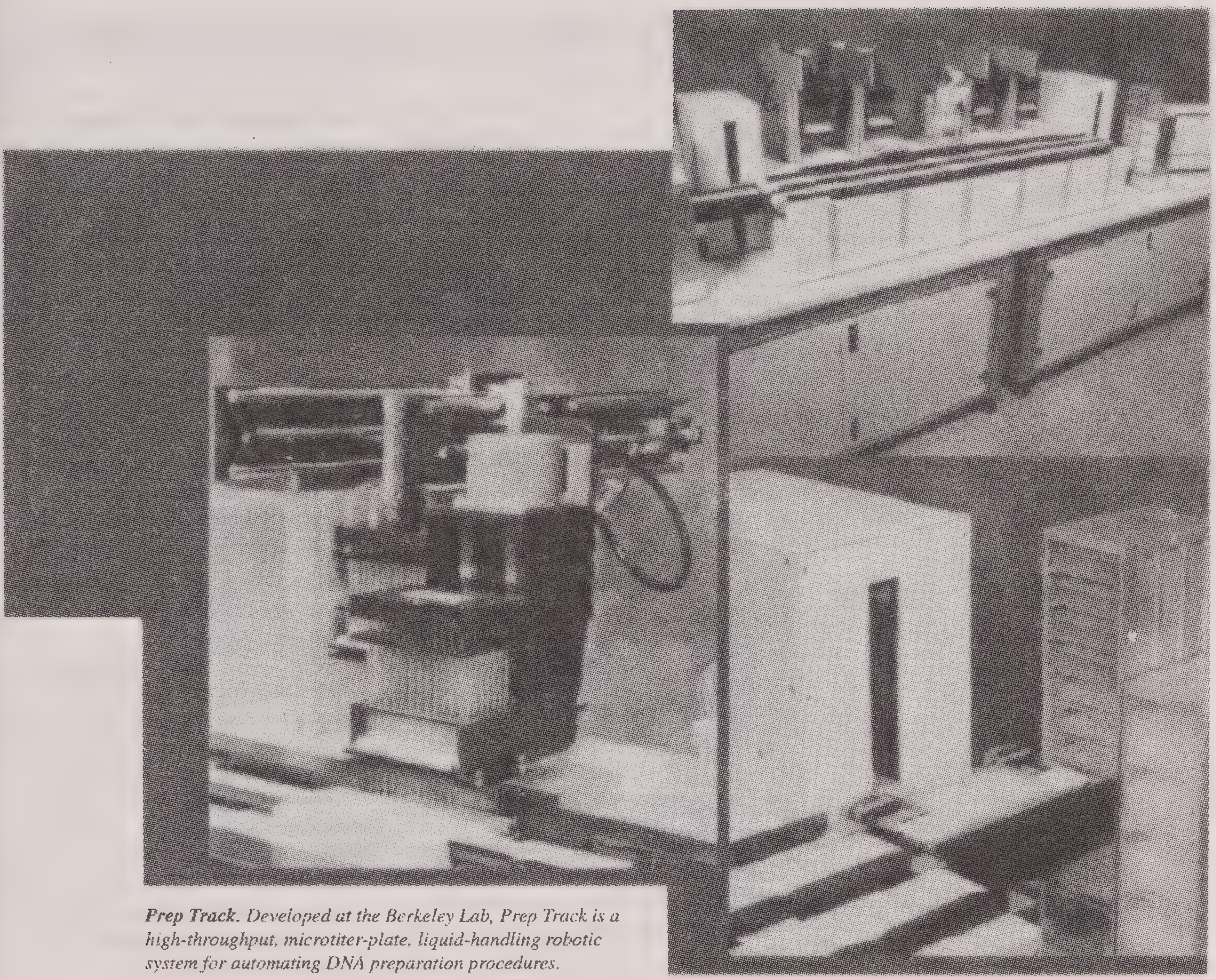
Berkeley Lab's near-term needs are for 960 samples per day of DNA extracted from overnight bacteria growths. The DNA protocol is a modified boil prep prepared in a 96-well format. Overnight bacteria growths are lysed, and samples are separated from cell debris by centrifugation. The DNA is recovered by ethanol precipitation.

Informatics

The informatics group is focused on hardware and software support and system administration, software development for end sequencing, transposon mapping and sequence template selection, data-flow automation, gene finding, and sequence analysis. Data-flow automation is the main emphasis. Six key steps have been identified in this process, and software is being written and tested to automate all six. The first step involves controlling gel quality, trimming vector sequence, and storing the sequences in a database. A program module called Move-Track-Trim, which is now used in production, was written to handle these steps. The second through fourth steps in this process involve assembling, editing, and reconstructing P1 clones of 80,000 base pairs from 400-base traces. The fifth step is sequence annotation, and the sixth is data submission.

Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, and gene signals such as promoters. LBNL is developing a workbench for automatic sequence annotation and annotation viewing and editing. The goal is to run a series of sequence-analysis tools and display the results to compare the various predictions. Researchers then will be able to examine all the annotations (for example, genes predicted by various gene-finding methods) and select the ones that look best.

Nomi Harris developed Genotator, an annotation workbench consisting of a stand-alone annotation browser and several sequence-analysis functions. The back end runs several gene finders, homology searches (using BLAST), and signal searches and saves the results in ".ace" format. Genotator thus automates the tedious process of operating a dozen different sequence-analysis programs with many different input and output formats. Genotator can function via command-line arguments or with the graphical user interface (*http://www-hgc.lbl.gov/inf/annotation.html*).



Prep Track. *Developed at the Berkeley Lab, Prep Track is a high-throughput, microtiter-plate, liquid-handling robotic system for automating DNA preparation procedures. Microtiter plates are fetched from cassettes, moved to one of two conveyor belts, and transported to protocol-defined modules. Plates are moved continuously and automatically through the system as each module simultaneously processes plates in the module lift stations. The plates exit the system and are stored in microtiter-plate cassettes.*

Modules include a station capable of dispensing liquids in volumes from as low as 5 microliters to several milliliters, four 96-channel pipettors, and the plate-fetching module. Each module is controlled independently by programmable logic controllers (PLCs). The overall system is controlled by a personal computer and a Visual Basic Control master that determines the order in which plates are processed. The actions of each lift station and dispenser or pipettor are determined locally by programs resident in each module's PLC. The Visual Basic Control program moves the plates through the system based on the predefined protocol and on module status reports as monitored by PLCs.

The current belt length on the Prep Track supports eight standard modules, which can be reconfigured to any order. Standardization of mechanical, electrical, and communication components allows new modules to be designed and manufactured easily. The current standard module footprint is 250 mm wide, 600 mm deep, and 250 mm to the conveyor belt deck. The first protocol to be implemented on Prep Track will be polymerase chain reaction setups, with sequence-reaction setups to follow. [Source: LBNL]

Progress to Date

Chromosome 5

Over the last year, the center has focused its production genomic sequencing on the distal 40 megabases of the human chromosome 5 long arm. This region was chosen because it contains a cluster of growth factor and receptor genes and is likely to yield new and functionally related genes through long-range sequence analysis. Results to date include:

- 40-megabase nonchimeric map containing 82 yeast artificial chromosomes (YACs) in the chromosome 5 distal long arm.
- 20-megabase contig map in the region of 5q23-q33 that contains 198 P1s, 60 P1 artificial chromosomes, and 495 bacterial artificial chromosomes (BACs) linked by 563 sequenced tagged sites (STSs) to form contigs.
- 20-megabase bins containing 370 BACs in 74 bins in the region of 5q33-q35.

Chromosome 21

An early project in the study of Down syndrome (DS), which is characterized by chromosome 21 trisomy, constructed a high-resolution clone map in the chromosome 21 DS region to be used as a pilot study in generating a contiguous gene map for all of chromosome 21. This project has integrated P1 mapping efforts with transgenic studies in the Life Sciences Division. P1 maps provide a suitable form of genomic DNA for isolating and mapping cDNA.

- 186 clones isolated in the major DS region of chromosome 21 comprising about 3 megabases of genomic DNA extending from D21S17 to ETS2. Through cross-hybridization, overlapping P1s were identified, as well as gaps between two P1 contigs, and transgenic mice were created from P1 clones in the DS region for use in phenotypic studies.

Transgenic Mice

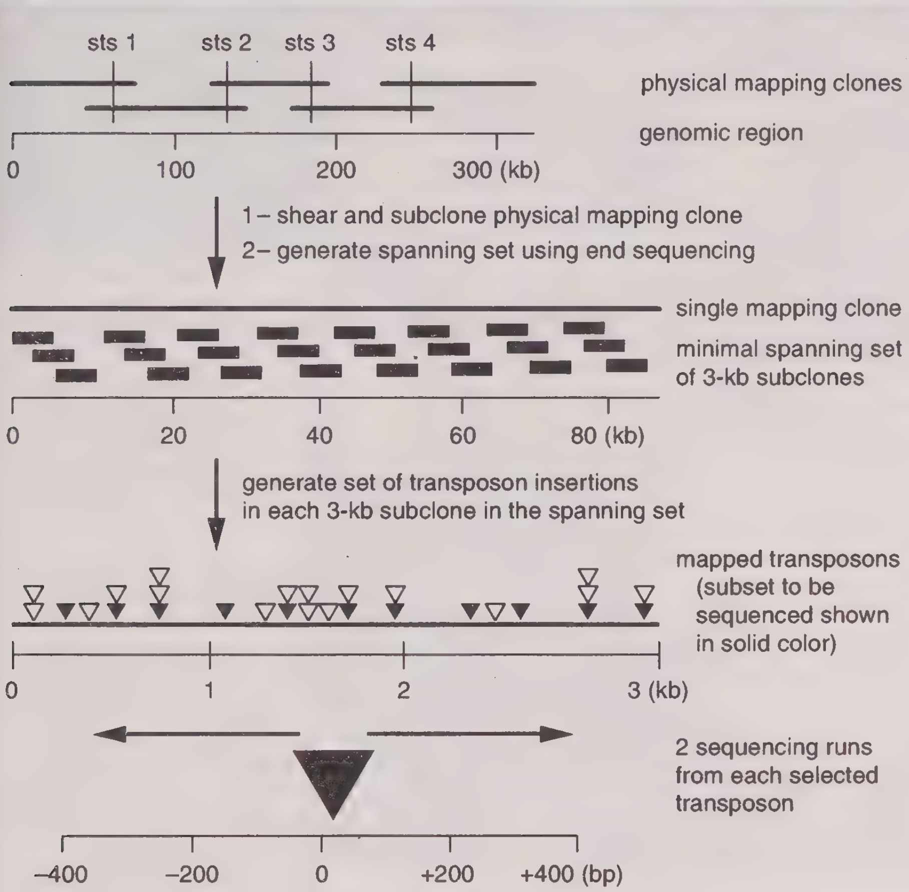
One of the approaches for determining the biological function of newly identified genes uses YAC transgenic mice. Human sequence harbored by YACs in transgenic mice has been shown to be correctly regulated both temporally and spatially. A set of nonchimeric overlapping YACs identified from the 5q31 region has been used to create transgenic mice. This set of transgenic mice, which together harbor 1.5 megabases of human sequence, will be used to assess the expression pattern and potential function of putative genes discovered in the 5q31 region. Additional mapping and sequencing are under way in a region of human chromosome 20 amplified in certain breast tumor cell lines.

Resource for Molecular Cytogenetics

Divining landmarks for human disease amid the enormous plain of the human genetic map is the mission of an ambitious partnership among the Berkeley Lab; University of California, San Francisco; and a diagnostics company. The collaborative Resource for Molecular Cytogenetics is charting a course toward important sites of biological interest on the 23 pairs of human chromosomes (*http://rmc-www.lbl.gov*).

The Resource employs the many tools of molecular cytogenetics. The most basic of these tools, and the cornerstone of the Resource's portfolio of proprietary technology, is a method generally known as "chromosome painting," which uses a technique referred to as fluorescence in situ hybridization or FISH. This technology was invented by LBNL Resource leaders Joe Gray and Dan Pinkel.

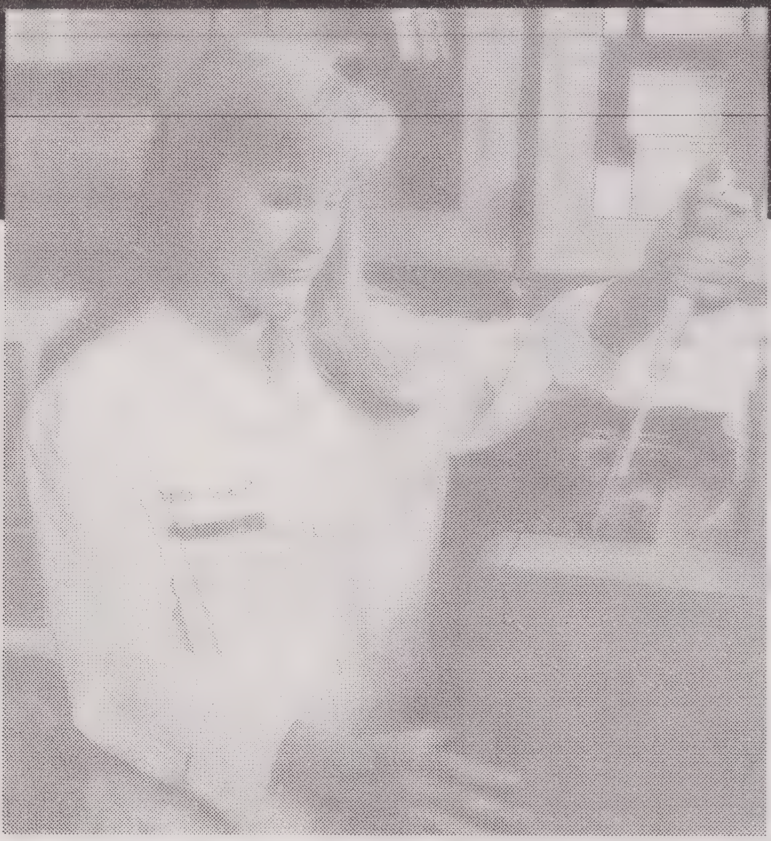
A technology to emerge recently from the Resource is known as "Quantitative DNA Fiber Mapping (QDFM)." High-resolution human genome maps in a form suitable for DNA sequencing traditionally have been constructed by


sts 1
sts 2
sts 3
sts 4
physical mapping clones
genomic region
0
100
200
300 (kb)
1– shear and subclone physical mapping clone
2– generate spanning set using end sequencing
single mapping clone
minimal spanning set of 3-kb subclones
0
20
40
60
80 (kb)
generate set of transposon insertions in each 3-kb subclone in the spanning set
mapped transposons (subset to be sequenced shown in solid color)
0
1
2
3 (kb)
2 sequencing runs from each selected transposon
–400
–200
0
+200
+400 (bp)

Sequencing Strategy. *The directed sequencing strategy used at LBNL involves four steps: (1) generate a P1-based physical map (using STS-content mapping) to provide a set of minimally overlapping clones, (2) shear and subclone each P1 clone into 3-kilobase fragments and identify a minimally overlapping subclone set, (3) generate and map transposon inserts in each subclone, and (4) sequence using commercial primer-binding sites engineered into the transposon. Subclone sequences are then assembled and edited, and the gaps are identified. P1 clones are reconstructed, and the resulting composite data is analyzed, annotated, and finally submitted to the databases. The production sequencing effort has generated 12 megabases of finished, double-stranded genomic DNA sequence from both* Drosophila *and human templates.* *[Source: Adapted from figure provided by LBNL]*

various methods of fingerprinting, hybridization, and identification of overlapping STSs. However, these techniques do not readily yield information about sequence orientation, the extent of overlap of these elements, or the size of gaps in the map. Ulli Weier of the Resource developed the QDFM method of physical map assembly that enables the mapping of cloned DNA directly onto linear, fully extended DNA molecules. QDFM allows unambiguous assembly of critical elements leading to high-resolution physical maps. This task now can be accomplished in less than 2 days, as compared with weeks by conventional methods. QDFM also enables detection and characterization of gaps in existing physical maps—a crucial step toward completing a definitive human genome map.

Lawrence Livermore National Laboratory scientist Stephanie Stilwagen loads a sample into an automated DNA sequencing system. [Source: Linda Ashworth, LLNL]



Research Narratives

University of Washington Genome Center

http://www.genome.washington.edu

The Human Genome Project soon will need to increase rapidly the scale at which human DNA is analyzed. The ultimate goal is to determine the order of the 3 billion bases that encode all heritable information. During the 20 years since effective methods were introduced to carry out DNA sequencing by biochemical analysis of recombinant-DNA molecules, these techniques have improved dramatically. In the late 1970s, segments of DNA spanning a few thousand bases challenged the capacity of world-class sequencing laboratories. Now, a few million base pairs per year represent state-of-the-art output for a single sequencing center.

However, the Human Genome Project is directed toward completing the human sequence in 5 to 10 years, so the data must be acquired with technology available now. This goal, while clearly feasible, poses substantial organizational and technical challenges. Organizationally, genome centers must begin building data-production units capable of sustained, cost-effective operation. Technically, many incremental refinements of current technology must be introduced, particularly those that remove impediments to increasing the scale of DNA sequencing. The University of Washington (UW) Genome Center is active in both areas.

Production Sequencing

Both to gain experience in the production of high-quality, low-cost DNA sequence and to generate data of immediate biological interest, the center is sequencing several regions of human and mouse DNA at a current throughput of 2 million bases per year. This "production sequencing" has three major targets: the human leukocyte antigen (HLA) locus on human chromosome 6, the mouse locus encoding the alpha subunit of T-cell receptors, and an "anonymous" region of human chromosome 7.

The HLA locus encodes genes that must be closely matched between organ donors and organ recipients. This sequence data is expected to lead to long-term improvements in the ability to achieve good matches between unrelated organ donors and recipients.

The mouse locus that encodes components of the T-cell–receptor family is of interest for several reasons. The locus specifies a set of proteins that play a critical role in cell-mediated immune responses. It provides sequence data that will help in the design of new experimental approaches to the study of immunity in mice—one of the most important experimental animals for immunological research. In addition, the locus will provide one of the first large blocks of DNA sequence for which both human and mouse versions are known.

Human-mouse sequence comparisons provide a powerful means of identifying the most important biological features of DNA sequence because these features are often highly conserved, even between such biologically different organisms as human and mouse. Finally, sequencing an "anonymous" region of human chromosome 7, a region about which little was known previously, provides experience in carrying out large-scale sequencing under the conditions that will prevail throughout most of the Human Genome Project.

Technology for Large-Scale Sequencing

In addition to these pilot projects, the UW Genome Center is developing incremental improvements in current sequencing technology. A particular focus is on enhanced computer software to process raw data acquired with automated laboratory instruments that are used in DNA mapping and sequencing. Advanced instrumentation is commercially available for determining DNA sequence via the "four-color–fluorescence method," and this instrumentation is expected to carry

University of Washington
Genome Center
Department of Medicine
Box 352145
Seattle, WA 98195

Maynard Olson
Director
206/685-7366, Fax: -7344
mvo@u.washington.edu

For more information on research projects and investigators at the University of Washington Genome Center, see abstracts in Part 2 of this report and the center's Web site (see URL above).

the main experimental load of the Human Genome Project. Raw data produced by these instruments, however, require extensive processing before they are ready for biological analysis.

Large-scale sequencing involves a "divide-and-conquer" strategy in which the huge DNA molecules present in human cells are broken into smaller pieces that can be propagated by recombinant-DNA methods. Individual analyses ultimately are carried out on segments of less than 1000 bases. Many such analyses, each of which still contains numerous errors, must be melded together to obtain finished sequence. During the melding, errors in individual analyses must be recognized and corrected. In typical large-scale sequencing projects, the results of thousands of analyses are melded to produce highly accurate sequence (less than one error in 10,000 bases) that is continuous in blocks of 100,000 or more bases. The UW Genome Center is playing a major role in developing software that allows this process to be carried out automatically with little need for expert intervention. Software developed in the UW center is used in more than 50 sequencing laboratories around the world, including most of the large-scale sequencing centers producing data for the Human Genome Project.

High-Resolution Physical Mapping

The UW Genome Center also is developing improved software that addresses a higher-level problem in large-scale sequencing. The starting point for large-scale sequencing typically is a recombinant-DNA molecule that allows propagation of a particular human genomic segment spanning 50,000 to 200,000 bases. Much effort during the last decade has gone into the physical mapping of such molecules, a process that allows huge regions of chromosomes to be defined in terms of sets of overlapping recombinant-DNA molecules whose precise positions along the chromosome are known. However, the precision required for knowing relationships of recombinant-DNA molecules derived from neighboring chromosomal portions increases as the Human Genome Project shifts its emphasis from mapping to sequencing.

High-resolution maps both guide the orderly sequencing of chromosomes and play a critical role in quality control. Only by mapping recombinant-DNA molecules at high resolution can subtle defects in particular molecules be recognized. Such defective human DNA sources, which are not faithful replicas of the human genome, must be weeded out before sequencing can begin. The UW Genome Center has a major program in high-resolution physical mapping which, like the work on sequencing itself, uses advanced computing tools. The center is producing maps of regions targeted for sequencing on a just-in-time basis. These highly detailed maps are proving extremely valuable in facilitating the production of high-quality sequence.

Ultimate Goal

Although many challenges currently posed by the Human Genome Project are highly technical, the ultimate goal is biological. The project will deliver immense amounts of high-quality, continuous DNA sequence into publicly accessible databases. These data will be annotated so that biologists who use them will know the most likely positions of genes and have convenient access to the best available clues about the probable function of these genes. The better the technical solutions to current challenges, the better the center will be able to serve future users of the human genome sequence.

Genome Database

http://www.gdb.org

The release of Version 6 of the Genome Database (GDB) in January 1996 signaled a major change for both the scientific community and GDB staff. GDB 6.0 introduced a number of significant improvements over previous versions of GDB, most notably a revised data representation for genes and genomic maps and a new curatorial model for the database. These new features, along with a remodeled database structure and new schema and user interface, provide a resource with the potential to integrate all scientific information currently available on human genomics. GDB rapidly is becoming the international biomedical research community's central source for information about genomic structure, content, diversity, and evolution.

A New Data Model

Inherent in the underlying organization of information in GDB is an improved model for genes, maps, and other classes of data. In particular, genomic segments (any named region of the genome) and maps are being expanded regularly. New segment types have been added to support the integration of mapping and sequencing data (for example, gene elements and repeats) and the construction of comparative maps (syntenic regions). New map types include comparative maps for representing conserved syntenies between species and comprehensive maps that combine data from all the various submitted maps within GDB to provide a single integrated view of the genome. Experimental observations such as order, size, distance, and chimerism are also available.

Through the World Wide Web, GDB links its stored data with many other biological resources on the Internet. GDB's External Link category is a growing collection of cross-references established between GDB entities and related information in other databases. By providing a place for these cross-references, GDB can serve as a central point of inquiry into technical data regarding human genomics.

Direct Community Data Submission and Curation

Two methods for data submission are in use. For individuals submitting small amounts of data, interactive editing of the database through the Web became available in April 1996, and the process has undergone several simplifications since that time. This continues to be an area of development for GDB because all editing must take place at the Baltimore site, and Internet connections from outside North America may be too slow for interactive editing to be practical. Until these difficulties are resolved, GDB encourages scientists with limited connectivity to Baltimore to submit their data via more traditional means (e-mail, fax, mail, phone) or to prepare electronic submissions for entry by the data group on site.

For centers submitting large quantities of data, GDB developed an electronic data submission (EDS) tool, which provides the means to specify login password validation and commands for inserting and updating data in GDB. The EDS syntax includes a mechanism for relating a center's local naming conventions to GDB objects. Data submitted to GDB may be stored privately for up to 6 months before it automatically becomes public. The database is programmed to enforce this Human Genome Project policy. Detailed specifications of GDB's EDS syntax and other submission instructions are available (EDS prototype, *http://www.gdb.org/eds*).

Since the EDS system was implemented, GDB has put forth an aggressive effort to increase the amount of data stored in the database. Consequently, the database has grown tremendously. During 1996 it grew from 1.8 to 6.7 gigabytes.

To provide accountability regarding data quality, the shift to community curation introduced the idea that individuals and

Genome Database
Johns Hopkins University
2024 E. Monument Street
Baltimore, MD 21205-2236

Stanley Letovsky
Informatics Director

Robert Cottingham
Operations Director

Telephone for both: 410/955-9705
Fax for both: 410/614-0434

David Kingsbury
Director, 1993–97*

In lieu of individual abstracts, research projects and investigators at GDB are represented in this narrative. More information can be found on GDB's Web site (see URL above).

---

*Now at Chiron Pharmaceuticals, Emeryville, California laboratories own the data they submit to GDB and that other researchers cannot modify it. However, others should be able to add information and comments, so an additional feature is the community's ability to conduct electronic online public discussions by annotating the database submissions of fellow researchers. GDB is the first database of its kind to offer this feature, and the number of third-party annotations is increasing in the form of editorial commentary, links to literature citations, and links to other databases external to GDB. These links are an important part of the curatorial process because they make other data collections available to GDB users in an appropriate context.

Improved Map Representation and Querying

Accompanying the release of GDB 6.0, the program Mapview creates graphical displays of maps. Mapview was developed at GDB to display a number of map types (cytogenetic, radiation hybrid, contig, and linkage) using common graphical conventions found in the literature. Mapview is designed to stand alone or to be used in conjunction with a Web browser such as Netscape, thereby creating an interactive graphical display system. When used with Netscape, Mapview allows the user to retrieve details about any displayed map object.

Maps are accessed through the query form for genomic segment and its subclasses via a special program that allows the user to select whole maps or slices of maps from specific regions of interest and to query by map type. The ability to browse maps stored in GDB or download them in the background was also incorporated into GDB 6.0.

GDB stores many maps of each chromosome, generated by a variety of mapping methods. Users who are interested in a region, such as the neighborhood of a gene or marker, will be able to see all maps that have data in that region, whether or not they contain the desired marker. To support database querying by region of interest, integrated maps have been developed that combine data from all the maps for each chromosome. These are called *Comprehensive Maps*.

Queries for all loci in a region of interest are processed against the comprehensive maps, thereby searching all relevant maps. Comprehensive maps are also useful for display purposes because they organize the content of a region by class of locus (e.g., gene, marker, clone) rather than by data source. This approach yields a much less complex presentation than an alignment of numerous primary maps. Because such information as detailed orders, order discrepancies between maps, and nonlinear metric relations between maps is not always captured in the comprehensive maps, GDB continues to provide access to aligned displays of primary maps.

A Variety of Searching Strategies

Recognizing the eclectic user community's need to search data and formulate queries, GDB offers a spectrum of simple to complex search strategies. In addition, direct programming access is available using either GDB's object query language to the Object Broker software layer or standard query language to the underlying Sybase relational database.

Querying by Object Directly from GDB's Home Page

The simplest methods search for objects according to known GDB accession numbers; sequence database–accession numbers; specified names, including wildcard symbols that will automatically match synonyms and primary names; and keywords contained anywhere in the text.

Querying by Region of Interest

A region of interest can be specified using a pair of flanking markers, which can be cytogenetic bands, genes, amplimers (sequence tagged sites), or any other mapped objects. Given a region of interest, the comprehensive maps are searched to find all loci that fall within them. These loci can be displayed in a table, graphically as a slice through a comprehensive map, or as slices through a chosen set of primary maps. A comprehensive map slice shows all loci in the region, including genes, expressed sequence tags (ESTs), amplimers, and clones. A region also can be specified as a neighborhood around a single marker of interest.

Results of queries for genes, amplimers, ESTs, or clones can be displayed on a GDB comprehensive map. Results are spread across several chromosomes displayed in Mapview (see figure, p. 52). A query for all the PAX genes (specified as symbol = PAX* on the gene query form) retrieves genes on multiple chromosomes. Double-clicking on one of these genes brings up detailed gene information via the Web browser.

Querying by Polymorphism

GDB contains a large number of polymorphisms associated with genes and other markers. Queries can be constructed for a particular type of marker (e.g., gene, amplimer, clone), polymorphism (i.e., dinucleotide repeat), or level of heterozygosity. These queries can be combined with positional queries to find, for example, polymorphic amplimers in a region bounded by flanking markers or in a particular chromosomal band. If desired, the retrieved markers can be viewed on a comprehensive map.

Work in Progress

Mapview 2.3

Mapview 2.1, the next generation of the GDB map viewer, was released in March 1997. The latest version, Mapview 2.3, is available in all common computing environments because it is written in the Java programming language. Most important, the new viewer can display multiple aligned maps side by side in the window, with alignment lines indicating common markers in neighboring maps. As before, users can select individual markers to retrieve more information about them from the database.

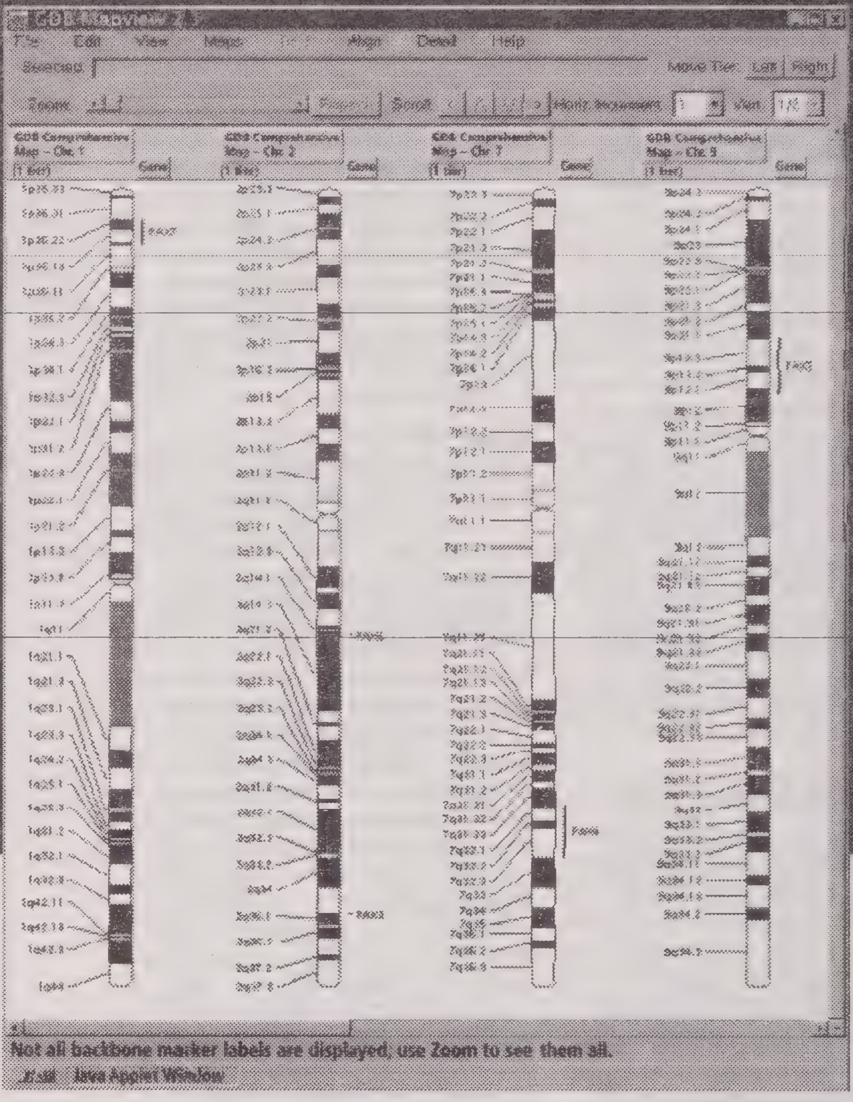
GDB developers have entered into a collaborative relationship with other members of the bioWidget Consortium so the Java-based alignment viewer will become part of a collection of freely available software tools for displaying biological data (*http://goodman.jax.org/projects/biowidgets/consortium*).

Future plans for Mapview include providing or enhancing the ability to generate manuscript-ready Postscript map images, highlight or modify the display of particular classes of map objects based on attribute values, and requery for additional information.

Variation

Since its inception, GDB has been a repository for polymorphism data, with more than 18,000 polymorphisms now in GDB. A collaboration has been initiated with the Human Gene Mutation Database (HGMD) based in Cardiff, Wales, and headed by David Cooper and Michael Krawczak. HGMD's extensive collection of human mutation data, covering many disease-causing loci, includes sequence-level mutation characterizations. This data set will be included in GDB and updated from HGMD on an ongoing basis. The HGMD team also will provide advice

Graphical Display of Results of Query for Genes with Names matching "PAX." (Source: Robert Cottingham, GDB)*

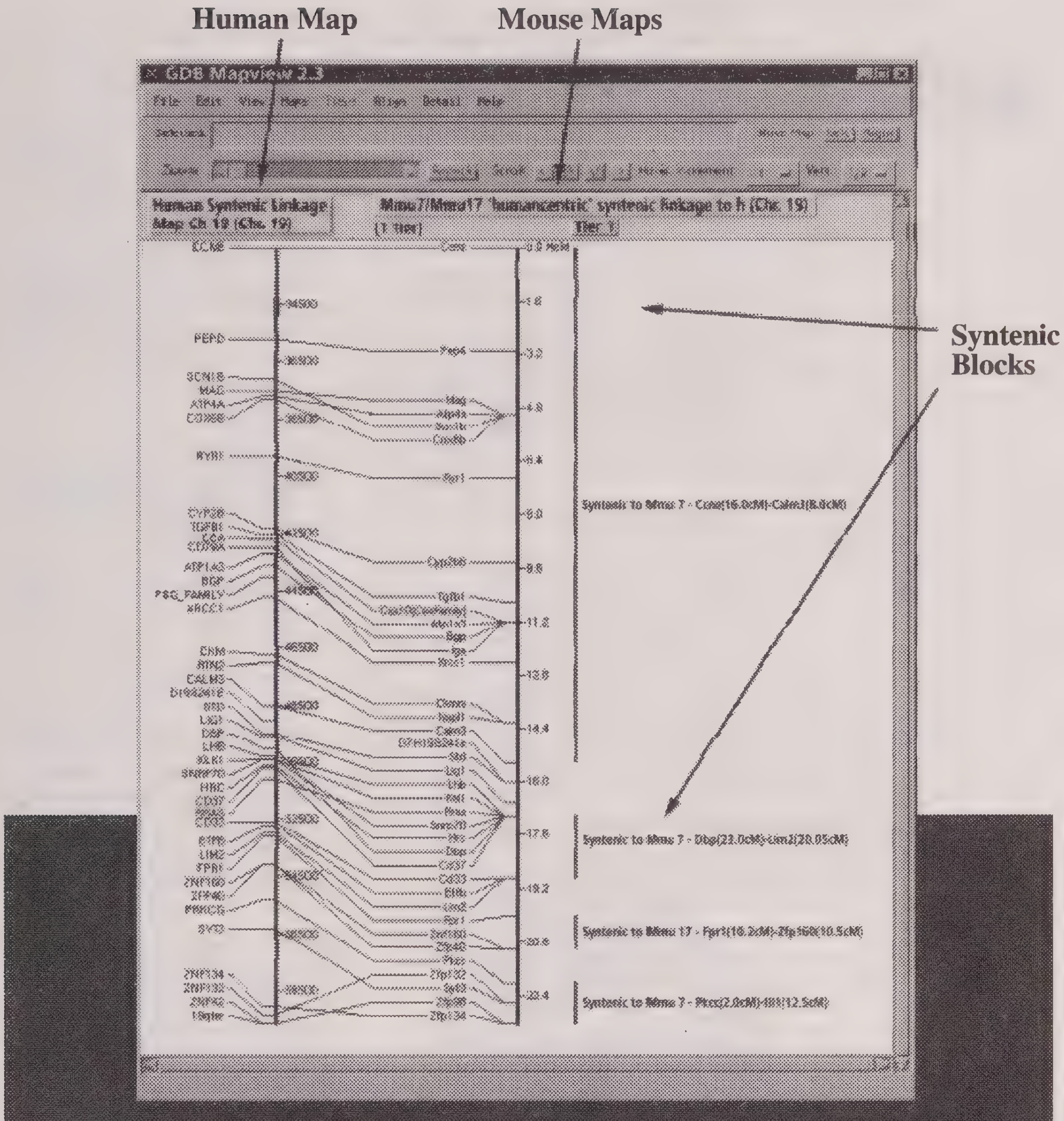

File
Edit
View
Maps
Align
Detail
Help
Zoom
Scroll
GDB Comprehensive Map – Chr 1
GDB Comprehensive Map – Chr 2
GDB Comprehensive Map – Chr 7
GDB Comprehensive Map – Chr 9
Gene
PAX7
PAX8
PAX3
PAX4
PAX5
Java Applet Window

Not all backbone marker labels are displayed, use Zoom to see them all.

on GDB's representation of genetic variation, which is being enhanced to model mutations and polymorphisms at the sequence level. These modifications will allow GDB to act as a repository for single-nucleotide polymorphisms, which are expected to be a major source of information on human genetic variation in the near future.

Mouse Synteny

Genomic relationships between mouse and man provide important clues regarding gene location, phenotype, and function (see figure, p. 53). One of GDB's goals is to enable direct comparisons between these two organisms, in collaboration with the Mouse Genome Database


Human Map
Mouse Maps
Syntenic Blocks
GDB Mapview 2.3
File
Edit
View
Maps
Align
Detail
Help
Zoom
Scroll
Human Syntenic Linkage Map Ch 19 (Chr. 19)
Mmu7/Mmu17 'humancentric' syntenic linkage to h (Chr. 19)
Tier 1
PEPD
SCN1B
MAG
ATP4A
RYR1
CYP2B
TGFB1
ATP1A3
BGP
PSG_FAMILY
XRCC1
CKM
CALM3
ZNF134
Syntenic to Mmu 7 - Ccne(16.0cM)-Calm3(8.0cM)
Syntenic to Mmu 7 - Dbp(22.0cM)-Lim2(28.05cM)
Syntenic to Mmu 17 - Fpr1(10.2cM)-Zfp160(10.5cM)
Syntenic to Mmu 7 - Ptcc(2.0cM)-(12.5cM)

Rearranged Mouse Map Aligned Against Human Chromosome (Source: Robert Cottingham) at Jackson Laboratory. GDB is making additions to its schema to represent this information so that it can be displayed graphically with Mapview. In addition, algorithmic work is under way to use mapping data to automatically identify regions of conserved synteny between mouse and man. These algorithms will allow the synteny maps to be updated regularly. An important application of comparative mapping is the ability to predict the existence and location of unknown human homologs of known, mapped mouse genes. A set of such predictions is available in a report at the GDB Web site, and similar data will be available in the database itself in the spring of 1998.

Collaborations

GDB is a participant in the Genome Annotation Consortium (GAC) project, whose goal is to produce high-quality, automatic annotation of genomic sequences (*http://compbio.ornl.gov/CoLab*). Currently, GDB is developing a prototype mechanism to transition from GDB's Mapview display to the GAC sequence-level browser over common genome regions. GAC also will establish a human genome reference sequence that will be the base against which GDB will refer all polymorphisms and mutations. Ultimately, every genomic object in GDB should be related to an appropriate region of the reference sequence.

Sequencing Progress

The sequencing status of genomic regions now can be recorded in GDB. Based on submissions to sequence databases, GAC will determine genomic regions that have been completed. GDB also will be collaborating with the European Bioinformatics Institute, in conjunction with the international Human Genome Organisation (HUGO), to maintain a single shared Human Sequence Index that will record commitments and status for sequencing clones or regions. As a result, the sequencing status of any region can be displayed alongside other GDB mapping data.

Outreach

The Genome Database continues to seek direct community feedback and interact with the broader science community via various sources:

- International Scientific Advisory Committee meets annually to offer input and advice.
- Quarterly Review Committee confers frequently with the staff to track GDB progress and suggest change.
- HUGO nomenclature, chromosome, and other editorial committees have specialized functions within GDB, providing official names and consensus maps and ensuring the high quality of GDB's content.

Copies of GDB are available worldwide from ten mirror sites (nodes) that make the data more easily accessible to the international research community. GDB staff meet annually with node managers to facilitate interaction and to benefit from other user perspectives.

The National Center for Genome Resources (NCGR) is a not-for-profit organization created to design, develop, support, and deliver resources in support of public and private genome and genetic research. To accomplish these goals, NCGR is developing and publishing the Genome Sequence DataBase (GSDB) and the Genetics and Public Issues (GPI) program.

NCGR is a center to facilitate the flow of information and resources from genome projects into both public and private sectors. A broadly based board of governors provides direction and strategy for the center's development.

NCGR opened in Santa Fe in July 1994, with its initial bioinformatics work being developed through a cooperative 5-year agreement with the Department of Energy funded in July 1995. Committed to serving as a resource for all genomic research, the center works collaboratively with researchers and seeks input from users to ensure that tools and projects under development meet their needs.

Genome Sequence DataBase

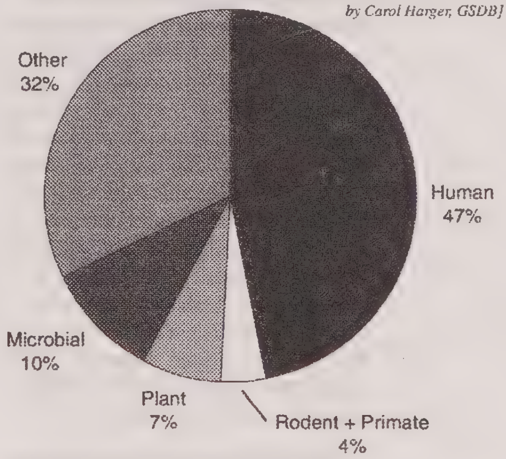
GSDB is a relational database that contains nucleotide sequence data (see pie chart) and its associated annotation from all known organisms (*http://www.ncgr.org/gsdb*). All data are freely available to the public. The major goals of GSDB are to provide the support structure for storing sequence data and to furnish useful data-retrieval services.

GSDB adheres to the philosophy that the database is a "community-owned" resource that should be simple to update to reflect new discoveries about sequences. A corollary to this is GSDB's conviction that researchers know their areas of expertise much better than a database curator and, therefore, they should be given ownership and control over the data they submit to the database. The true role of the GSDB staff is to help researchers submit data to and retrieve data from the database.

GSDB Enhancements

During 1996, GSDB underwent a major renovation to support new data types and concepts that are important to genomic research. Tables within the database were restructured, and new tables and data fields were added. Some key additions to GSDB include the support of data ownership, sequence alignments, and discontiguous sequences.

The concept of data ownership is a cornerstone to the functioning of the new GSDB. Every piece of data (e.g., sequence or feature) within the database is owned by the submitting researcher, and changes can be made only by the data owner or GSDB staff. This implementation of data ownership provides GSDB with the ability to support community (third-party) annotation—the addition of annotation to a sequence by other community researchers.



This chart illustrates the taxonomic distribution of the 1,076,481,102 base pairs in the Genome Sequence DataBase. About 47% of the base pairs and 58% of the total database records represent human sequences (August 1997). *[Source: Adapted from chart provided by Carol Harger, GSDB]*

http://www.ncgr.org

Genome Sequence DataBase
1800 Old Pecos Trail, Suite A
Santa Fe, NM 87505

Peter Schad
Vice-President, Bioinformatics and Biotechnology
505/995-4447, Fax: -4432
cnc@ncgr.org

Carol Harger
GSDB Manager
505/982-7840, Fax: -7690
cah@ncgr.org

In lieu of individual abstracts, research projects and investigators at NCGR are represented in this narrative. More information can be found on the center's Web site (see URL above).

A second enhancement of GSDB is the ability to store and represent sequence alignments. GSDB staff has been constructing alignments to several key sequences including the env and pol (reverse transcriptase) genes of the HIV genome, the complete chromosome VIII of Saccharomyces cerevisiae, and the complete genome of *Haemophilus influenzae*. These alignments are useful as possible sites of biological interest and for rapidly identifying differences between sequences.

A third key GSDB enhancement is the ability to represent known relationships of order and distance between separate individual pieces of sequence. These sets of sequences and their relative positions are grouped together as a single discontiguous sequence. Such a sequence may be as simple as two primers that define the ends of a sequence tagged site (STS), it may comprise all exons that are part of a single gene, or it may be as complex as the STS map for an entire chromosome.

GSDB staff has constructed discontiguous sequences for human chromosomes 1 through 22 and X that include markers from Massachusetts Institute of Technology–Whitehead Institute STS maps and from the Stanford Human Genome Center. The set of 2000 STS markers for chromosome X, which were mapped recently by Washington University at St. Louis, also have been added to chromosome X. About 50 genomic sequences have been added to the chromosome 22 map by determining their overlap with STS markers. Genomic sequences are being added to all the chromosomes as their overlap with the STS markers is determined. These discontiguous sequences can be retrieved easily and viewed via their sequence names using the GSDB Annotator. Sequence names follow the format of HUMCHR#MP, where # equals 1 through 22 or X.

GSDB staff also has utilized discontiguous sequences to construct maps for maize and rice. The maize discontiguous sequences were constructed using markers from the University of Missouri, Columbia. Markers for the rice discontiguous sequence were obtained from the Rice Genome Database at Cornell University and the Rice Genome Research Project in Japan.

New Tools

As a result of the major GSDB renovation, new tools were needed for submitting and accessing database data. Annotator was developed as a graphical interface that can be used to view, update, and submit sequence data *(http://www.ncgr.org/gsdb/beta.html)*. Maestro, a Web-based interface, was developed to assist researchers in data retrieval (http://www.ncgr.org/gsdb/maestrobeta.html). Although both these tools currently are available to researchers, GSDB is continuing development to add increased capabilities.

Annotator displays a sequence and its associated biological information as an image, with the scale of the image adjustable by the user. Additional information about the sequence or an associate biological feature can be obtained in a pop-up window. Annotator also allows a user to retrieve a sequence for review, edit existing data, or add annotation to the record. Sequences can be created using Annotator, and any sequences created or edited can be saved either to a local file for later review and further editing or saved directly to the database.

Correct database structures are important for storing data and providing the research community with tools for searching and retrieving data. GSDB is making a concerted effort to expand and improve these services. The first generation of the Maestro query tool is available from the GSDB Web pages. Maestro allows researchers to perform queries on 18 different fields, some of which are queryable only through GSDB, for example, D segment numbers from the Genome Database at Johns Hopkins University in Baltimore.

Additionally, Maestro allows queries with mixed Boolean operators for a more refined search. For example, a user may wish to compare relatively long mouse and human sequences that do not contain identified coding regions. To obtain all sequences meeting these criteria, the scientific name field would be searched first for "Mus musculus" and then for "Homo sapiens" using the Boolean term "OR." Then the sequence-length filter could be used to refine the search to sequences longer than 10,000 base pairs. To exclude sequences containing identified coding-region features, the "BUT NOT" term can be used with the Feature query field set equal to "coding region."

With Maestro, users can view the list of search matches a few at a time and retrieve more of the list as needed. From the list, users can select one or several sequences according to their short descriptions and review or download the sequence information in GIO, FASTA, or GSDB flatfile format.

Future Plans

Although most pieces necessary for operation are now in place, GSDB is still improving functionality and adding enhancements. During the next year GSDB, in collaboration with other researchers, anticipates creating more discontiguous sequence maps for several model organisms, adding more functionality to and providing a Web-based submission tool and tool kit for creating GIO files.

Microbial Genome Web Pages

NCGR also maintains informational Web pages on microbial genomes. These pages, created as a community reference, contain a list of current or completed eubacterial, Archaeal, and eukaryotic genome sequencing projects. Each main page includes the name of the organism being sequenced, sequencing groups involved, background information on the organism, and its current location on the Carl Woese Tree of Life. As the Microbial Genome Project progresses, the pages will be updated as appropriate.

Genetics and Public Issues Program

GPI serves as a crucial resource for people seeking information and making decisions about genetics or genomics (http://www.ncgr.org/gpi). GPI develops and provides information that explains the ethical, legal, policy, and social relevance of genetic discoveries and applications.

To achieve its mission, GPI has set forth three goals: (1) preparation and development of resources, including careful delineation of ethical, legal, policy, and social issues in genetics and genomics; (2) dissemination of genetic information targeted to the public, legal and health professionals, policymakers, and decision makers; and (3) creation of an information network to facilitate interaction among groups.

GPI delivers information through four primary vehicles: online resources, conferences, publications, and educational programs. The GPI program maintains a continually evolving World Wide Web site containing a range of material freely accessible over the Internet.

Los Alamos National Laboratory researcher David Bruce uses an automated system for gridding chromosome library clones in preparation of very dense filter arrays for hybridization experiments. [Source: Lynn Clark, LANL]

Program Management

The Human Genome Program was conceived in 1986 as an initiative within the DOE Office of Health and Environmental Research, which has been renamed Office of Biological and Environmental Research (OBER) (see chart below). The program is administered primarily through the OBER Health Effects and Life Sciences Research Division (HELSRD), both directed by David A. Smith until his retirement in January 1996. Marvin Frazier is now Director of HELSRD, and OBER is led by Associate Director Aristides Patrinos, who also serves as Human Genome Program manager. Previous directors and managers are listed in the table below. OBER is within the Office of Energy Research, directed by Martha Krebs.

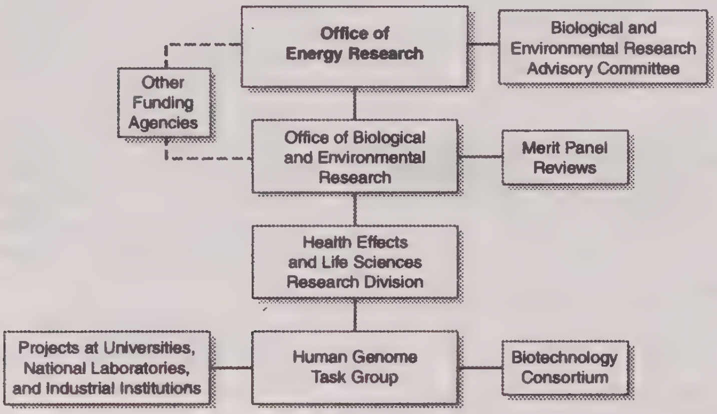
DOE OBER Mission

Based on mandates from Congress, DOE OBER's principal missions are to (1) develop the knowledge necessary to identify, understand, and anticipate long-term health and environmental consequences of energy use and development and (2) employ DOE's unique scientific and technological capabilities in solving major scientific problems in medicine, biology, and the environment.

Genome integrity and radiation biology have been a long-term concern of OBER at DOE and its predecessors—the Atomic Energy Commission (AEC) and the Energy Research and Development Administration (ERDA). In the United States, the first federal support

http://www.er.doe.gov/production/ober/hug_top.html

See Appendix A, p. 73, for information on Human Genome Project history, including enabling legislation.


Office of Energy Research
Biological and Environmental Research Advisory Committee
Other Funding Agencies
Office of Biological and Environmental Research
Merit Panel Reviews
Health Effects and Life Sciences Research Division
Projects at Universities, National Laboratories, and Industrial Institutions
Human Genome Task Group
Biotechnology Consortium

OHER Associate or Acting Directors		Human Genome Program Managers	
Charles De Lisi	1985	Benjamin J. Barnhart	1988
Robert W. Wood	1987	David A. Smith	1991
David J. Galas	1990	Aristides Patrinos	1996
Aristides Patrinos	1993		

Institutions Conducting DOE-Sponsored Genome Research	
DOE national laboratories	7
Academic institutions	28
Private-sector institutions	10
Companies, including Small Business Innovation Research	11
Foreign institutions (Russia, Canada, Israel)	7

for genetic research was through AEC. In the early days of nuclear energy development, the focus was on radiation effects and broadened later under ERDA and DOE to include health implications of all energy technologies and their by-products.

Today, extensive OBER-sponsored research programs on genomic structure, maintenance, damage, and repair continue at the national laboratories and universities. These and other OBER efforts support a DOE shift toward a preventive approach to health, environment, and safety concerns. World-class scientists in top facilities working on leading-edge problems spawn the knowledge to revolutionize the technology, drive the future, and add value to the U.S. economy. Major OBER research includes characterization of DNA repair genes and improvement of methodologies and resources for quantifying and characterizing genetic polymorphisms and their relationship to genetic susceptibilities.

To carry out its national research and development obligations, OBER conducts the following activities:

- Sponsors peer-reviewed research and development projects at universities, in the private sector, and at DOE national laboratories (see box, p. 59).
- Considers novel, beneficial initiatives with input from the scientific community and governmental sectors.
- Provides expertise to various governmental working groups.
- Supports the capabilities of multidisciplinary DOE national laboratories and their unique user facilities for the nation's benefit (p. 61).

Human Genome Program resources and technologies are focused on sequencing the human genome and related informatics and supportive infrastructure (see chart and tables, p. 62). The genomes of selected microorganisms are analyzed under the separate Microbial Genome Program.

DOE Human Genome Task Group

Member	Specialty
Chair: Aristides Patrinos	Physical sciences
Benjamin J. Barnhart	Genetics, Radiation biology
Elbert Branscomb	Scientific Director, Joint Genome Institute
Daniel W. Drell	Biology, ELSI, Informatics, Microbial genome
Ludwig Feinendegen	Medicine, Radiation biology
Marvin Frazier	Molecular and cellular biology
Gerald Goldstein†	Physical science, Instrumentation
D. Jay Grimes†	Microbiology
Roland Hirsch	Structural biology, Instrumentation
Arthur Katz*	Physical sciences
Anna Palmisano*†	Microbiology, Microbial genome
Michael Riches	Physical sciences
Jay Snoddy†	Molecular biology, Informatics
Marvin Stodolsky	Molecular biology, Biophysics
David G. Thomassen	Cell and molecular biology
John C. Wooley	Computational biology

---

*Joined, 1997.
†Left OBER, 1997.

Biotechnology Consortium

Chair: Aristides Patrinos	DOE Office of Biological and Environmental Research
Charles Arntzen*	Cornell University
Elbert Branscomb	Lawrence Livermore National Laboratory
Charles Cantor	Boston University
Anthony Carrano	Lawrence Livermore National Laboratory
Thomas Caskey	Merck Research Laboratories
David Eisenberg	University of California, Los Angeles
Chris Fields†	National Center for Genome Resources
David Galas	Darwin Molecular, Inc.
Raymond Gesteland	University of Utah
Keith Hodgson	Stanford University
Leroy Hood	University of Washington, Seattle
David Kingsbury†	Chiron Pharmaceuticals
Robert Moyzis†	University of California, Irvine
Mohandas Narla*	Lawrence Berkeley National Laboratory
Michael Palazzolo	Amgen, Inc.
Melvin Simon*	California Institute of Technology
Hamilton Smith*	Johns Hopkins University School of Medicine
Lloyd Smith	University of Wisconsin, Madison
Lisa Stubbs	Lawrence Livermore National Laboratory
Edward Uberbacher*	Oak Ridge National Laboratory
Marc Van Montagu*	Ghent University, Belgium
Executive Officer: Sylvia Spengler	Lawrence Berkeley National Laboratory

---

*Appointed after October 1996.
†Resigned, 1997.
Note: All members of the DOE Human Genome Task Group are ex-officio members of the Biotechnology Consortium.

Major DOE User Facilities and Resources Relevant to Molecular Biology Research

Although the genome program is contributing fundamental information about the structure of chromosomes and genes, other types of knowledge are required to understand how genes and their products function. Three-dimensional protein structure studies are still essential because structure cannot be predicted fully from its encoded DNA sequence.

To enhance these and other studies, DOE builds and maintains structural biology user facilities that enable scientists to gain an understanding of relationships between biological structures and their functions, study disease processes, develop new pharmaceuticals, and conduct basic research in molecular biology and environmental processes. These resources are used heavily by both academic and private-sector scientists.

Other important resources available to the research community include the clone libraries developed in the National Laboratory Gene Library Project and distributed worldwide, the GRAIL Online Sequence Interpretation Service, and the Mouse Genetics Research Facility.

Pacific Northwest National Laboratory
Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
Stanford Synchrotron Radiation Laboratory
Argonne National Laboratory
Brookhaven National Laboratory
Oak Ridge National Laboratory
Los Alamos National Laboratory

Argonne National Laboratory
  Advanced Photon Source

Brookhaven National Laboratory
  High-Flux Beam Reactor
  National Synchrotron Light Source
  Protein Structure Data Bank
  Scanning Transmission Electron Microscope

Lawrence Berkeley National Laboratory
  Advanced Light Source
  Center for X-Ray Optics
  National Energy Research Scientific Computing Center

Lawrence Livermore National Laboratory
  National Laboratory Gene Library Project

Los Alamos National Laboratory
  National Flow-Cytometry Resource
  National Laboratory Gene Library Project
  Neutron-Scattering Center

Oak Ridge National Laboratory
  GRAIL, Online Sequence Interpretation Service
  Mouse Genetics Research Facility

Pacific Northwest National Laboratory
  Environmental Molecular Sciences Laboratory

Stanford University
  Synchrotron Radiation Laboratory

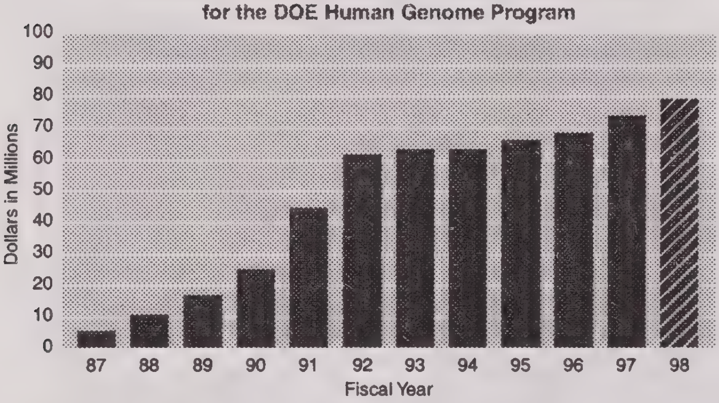
Human Genome Program

Coordination and Resources

Program coordination is the responsibility of the Human Genome Task Group (see box, p. 60), which, beginning in 1997, includes Elbert Branscomb, the Joint Genome Institute's Scientific Director. The task group is aided by the Biotechnology Consortium (which succeeded the former Human Genome Coordination Committee; see box, p. 60) to foster information exchange and dissemination. The task group administers the DOE Human Genome Program and its evolving needs and reports to the Associate Director for Biological and Environmental Research (currently Aristides Patrinos). The task group arranges periodic workshops and coordinates site reviews for genome centers, the Joint Genome Institute, databases, and other large projects. It also coordinates peer review of research proposals, administration of awards, and collaboration with all concerned agencies and organizations.

The Biotechnology Consortium provides the OBER Associate Director with external expertise in all aspects of genomics and informatics and a mechanism by which OBER can keep track of the latest developments in the field. It facilitates development and dissemination of novel genome technologies throughout the DOE system, ensures appropriate management and sharing of data and resources by all DOE contractors and grantees, and promotes interactions with other national and international genomic entities.

Operating Expenditures and FY 1998 Projected Budget for the DOE Human Genome Program



Human Genome Program Fiscal Year Expenditures ($M)

Year	Operating	Capital Equipment	Construction	Total
1996	68.3	5.6	5.7	79.6
1997	73.9	6.0	1.0	80.9
1998*	79.9	5.2	0.0	85.1

*Projected expenses.

Human Genome Program Operating Funds Distribution in FY 1996 ($K)

FY 1996	Mapping	Sequencing	Sequencing Technology	Informatics	ELSI	Administration	Totals	%
DOE Laboratories	8,980	11,015	11,128	6,840	313	2,783*	41,059	60.1
Academic	6,671	4,368	3,257	6,178	642	4	21,120	30.9
Nonprofit	563	0	467	2,783	1,311	38	5,162	7.5
Federal	0	0	0	0	0	1,000**	1,000	1.5
Total	16,214	15,383	14,852	15,801	2,266	3,825	68,341	
% of Total	23.8	22.5	21.7	23.1	3.3	5.6	100	

*Includes DOE laboratories' nonresearch costs but not U.S. government administration or SBIR.
**DOE contribution to the international Human Frontiers Neurosciences Program.

Communication

The DOE Human Genome Program communicates information in a variety of ways. These communication systems include the Human Genome Management Information System (HGMIS), projects in the Ethical, Legal, and Social Issues (ELSI) Program, electronic resources, meetings, and fellowships. Some of these mechanisms are described below. For more details, see Research Highlights, ELSI projects, p. 18.

HGMIS

HGMIS provides technical communication and information services for the DOE OBER Human Genome Program Task Group. HGMIS is charged with (1) helping to communicate genome-related matters and research to contractors, grantees, other (nongenome project) researchers, and other multipliers of information pertaining to genetic research; (2) serving as a clearinghouse for inquiries about the U.S. genome project; and (3) reducing research duplication by providing a forum for interdisciplinary information exchange (including resources developed) among genetic investigators worldwide.

HGMIS publishes the newsletter *Human Genome News*, sponsored by OBER. Over 14,000 *HGN* subscribers include genome and basic researchers at national laboratories, universities, and other research institutions; professors and teachers; industry representatives; legal personnel; ethicists; students; genetic counselors; physicians; science writers; and other interested individuals.

HGMIS also produces the DOE *Primer on Molecular Genetics*; a compilation of ELSI abstracts; and reports on the DOE Human Genome and Microbial Genome Programs, contractor-grantee workshops, and other related subjects.

Electronic versions of the primer and other HGMIS publications are available via the World Wide Web. HGMIS also initiates and maintains other related Web sites (see DOE Electronic Genome Resources section below and DOE Web Sites at right).

In addition to their print and online publishing efforts, HGMIS staff members answer questions generated via Web sites, telephone, fax, and e-mail. They also furnish customized information about the genome project for multipliers of information (contact: Betty Mansfield at 423/576-6669, Fax: /574-9888, *mansfieldbk@ornl.gov*).

DOE Electronic Genome Resources

Web Sites. The DOE Human Genome Program Home Page displays pointers to other programs within OBER and the Office of Energy Research. Links are made to additional biological and environmental information and to HGMIS, Genome Database, and other sites.

HGMIS initiates and maintains the searchable Human Genome Project Information Web site. This site contains more than 1700 text files of information for multidisciplinary technical audiences as well as for lay persons interested in learning about the science, goals, progress, and history of the project. Users include almost all levels of students; education, medical, and legal professionals; genetic society and support group members; biotechnology and pharmaceutical industry personnel; administrators; policymakers; and the press.

The site also houses a section of frequently asked questions, a quick fact finder, *Primer on Molecular Genetics*, all issues of *Human Genome News*, DOE Human Genome Program and contractor-grantee workshop reports, *To Know Ourselves*, historical documents, research abstracts, calendars of genome events, and hundreds of links to genome research and educational sites. More than 1000 other Web pages link to this site, resulting in more than 100,000 text file transfers each month. This

DOE Web Sites

DOE Human Genome Program
http://www.er.doe.gov/production/ober/hug_top.html

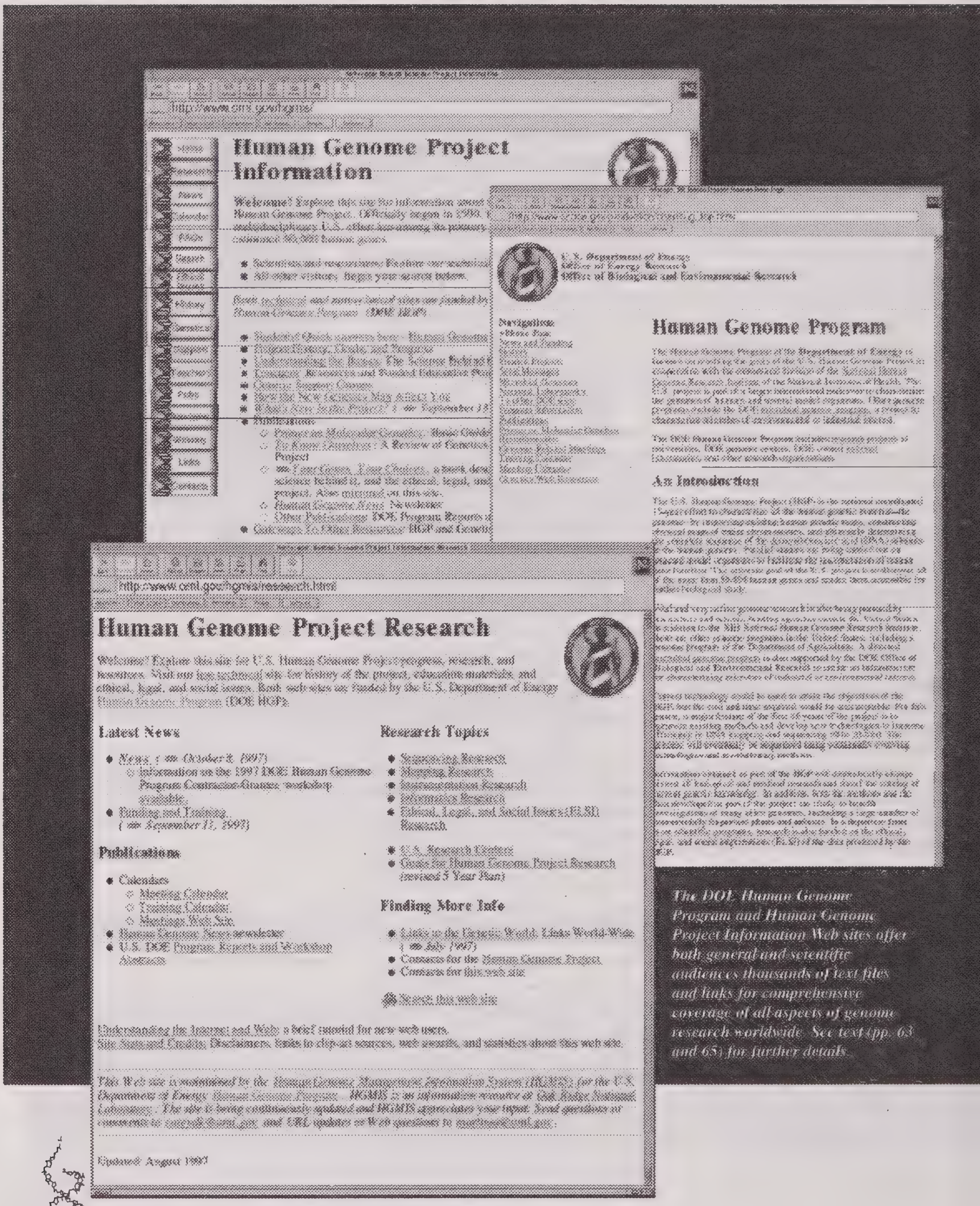
OBER
http://www.er.doe.gov/production/ober/ober_top.html

Office of Energy Research
http://www.er.doe.gov

Human Genome Project Information
http://www.ornl.gov/hgmis

HGP and Related Meetings
http://www.ornl.gov/meetings

Courts
http://www.ornl.gov/courts



http://www.ornl.gov/hgmis/

Human Genome Project Information

http://www.ornl.gov/hgmis/research.html

Human Genome Project Research

Welcome! Explore this site for U.S. Human Genome Project progress, research, and resources. Visit our less technical site for history of the project, education materials, and ethical, legal, and social issues. Both web sites are funded by the U.S. Department of Energy Human Genome Program (DOE HGP).

Latest News

- News (October 8, 1997)
  - Information on the 1997 DOE Human Genome Program Contractor-Grantee workshop available.
- Funding and Training (September 11, 1997)

Publications

- Calendars
  - Meeting Calendar
  - Training Calendar
  - Meetings Web Site
- Human Genome News newsletter
- U.S. DOE Program Reports and Workshop Abstracts

Research Topics

- Sequencing Research
- Mapping Research
- Instrumentation Research
- Informatics Research
- Ethical, Legal, and Social Issues (ELSI) Research

- U.S. Research Centers
- Goals for Human Genome Project Research (revised 5 Year Plan)

Finding More Info

- Links to the Genetic World: Links World-Wide (July 1997)
- Contacts for the Human Genome Project
- Contacts for this web site

Search this web site

Updated: August 1997

U.S. Department of Energy
Office of Energy Research
Office of Biological and Environmental Research

Human Genome Program

An Introduction

The DOE Human Genome Program and Human Genome Project Information Web sites offer both general and scientific audiences thousands of text files and links for comprehensive coverage of all aspects of genome research worldwide. See text (pp. 63 and 65) for further details.

HGMIS site has received a Four-Star designation from the Magellan Group and the Editor's Choice Award from LookSmart.

Genome-project and related meetings are listed at a Web site (see box, p. 63), through which users can register and submit research abstracts. Another listed related site discusses issues at the critical intersection of genetics and the court system. This Web page is part of a project to educate and prepare the judiciary for the coming onslaught of cases involving genetic issues and data.

Newsgroup. The Human Genome Program Newsgroup operates through the BIOSCI electronic bulletin board network to allow researchers worldwide to communicate, share ideas, and find solutions to problems. Genome-related information is distributed through the newsgroup, including requests for grant applications, reports from recent scientific and advisory meetings, announcements of future events, and listings of free software and services (*gnome-pr@net.bio.net* or *http://www.bio.net*).

Postdoctoral Fellowships

OBER established the Human Genome Distinguished Postdoctoral Research Program in 1990 to support research on projects related to the DOE Human Genome Program. Beginning in FY 1996, the Human Genome Distinguished Postdoctoral Fellowships were merged with the Alexander Hollaender Distinguished Postdoctoral Fellowships, which provide support in all areas of OBER-sponsored research. Postdoctoral programs are administered by the Oak Ridge Institute for Science and Education, a university consortium and DOE contractor. For additional information, contact Linda Holmes (423/576-3192, *holmesl@orau.gov*) or see the Web site (*http://www.orau.gov/ober/hollaend.htm*).

Human Genome Distinguished Postdoctoral Fellows

Names of past and current fellows in genome topics are given below with their research institutions and titles of proposed research. For 1996 research abstracts, refer to Index of Principal and Coinvestigators on p. 71 in Part 2 of this report.

1994 **Mark Graves** (Baylor College of Medicine): Graph Data Models for Genome Mapping

William Hawe (Duke University): Synthesis of Peptide Nucleic Acids for DNA Sequencing by Hybridization

Jingyue Ju (University of California, Berkeley): Design, Synthesis, and Use of Oligonucleotide Primers Labeled with Energy Transfer–Coupled Dyes

Mark Shannon (Oak Ridge National Laboratory): Comparative Study of a Conserved Zinc Finger Gene Region

1995 **Evan Eichler** (Lawrence Livermore National Laboratory): Identification, Organization, and Characterization of Zinc Finger Genes in a 2-Mb Cluster on 19p12

Kelly Ann Frazer (Lawrence Berkeley National Laboratory): In Vivo Complementation of the Murine Mutations Grizzled, Mocha, and Jitter

Soo-In Hwang (Lawrence Berkeley National Laboratory): Positional Cloning of Oncogenes on 20q13.2

James Labrenz (University of Washington, Seattle): Error Analysis of Principal Sequencing Data and Its Role in Process Optimization for Genome-Scale Sequencing Projects

Marie Ruiz-Martinez (Northeastern University): Multiplex Purification Schemes for DNA Sequencing–Reaction Products: Application to Gel-Filled Capillary Electrophoresis

Todd Smith (University of Washington, Seattle): Managing the Flow of Large-Scale DNA Sequence Information

Alexander Hollaender Distinguished Postdoctoral Fellows in Genome Research

1996 **Cymbeline Culiat** (Oak Ridge National Laboratory): Cloning of a Mouse Gene Causing Severe Deafness and Balance Defects

Tau-Mu Yi (Laboratory of Structural Biology and Molecular Medicine, Los Angeles): Structure-Function Analysis of Alpha-Factor Receptor

1997 **Jeffrey Koshi** (Los Alamos National Laboratory): Construction, Analysis, and Use of Optimal DNA Mutation Matrices

Sandra McCutchen-Maloney (Lawrence Livermore National Laboratory): Structure and Function of a Damage-Specific Endonuclease Complex

The laser-based flow cytometer developed at DOE national laboratories enables researchers to separate human chromosomes for analysis.
[Source: Los Alamos National Laboratory]

Coordination with Other Genome Programs

Enhancing genome research capabilities

The U.S. Human Genome Project is supported jointly by the Department of Energy (DOE) and the National Institutes of Health (NIH), each of which emphasizes different facets. The two agencies coordinate their efforts through development of common project goals and joint support of some programs addressing ethical, legal, and social issues (ELSI) arising from new genome tools, technology, and data.

Extraordinary advances in genome research are due to contributions by many investigators in this country and abroad. In the United States, such research (including nonhuman) also is funded by other federal agencies and private foundations and groups. Many countries are major contributors to the project through international collaborations and their own focused programs. Coordinating and facilitating these diverse research efforts around the world is the aim of the nongovernmental international Human Genome Organisation.

Some details of U.S. and worldwide coordination are provided below.

U.S. Human Genome Project: DOE and NIH

In 1988 DOE and NIH developed a Memorandum of Understanding that formalized the coordination of their efforts to decipher the human genome and thus "enhance the human genome research capabilities of both agencies." In early 1990 they presented Congress with a joint plan, *Understanding Our Genetic Inheritance, The U.S. Human Genome Project: The First Five Years (1991–1995)*. Referred to as the Five-Year Plan, it contained short-term scientific goals for the coordinated, multiyear research project and a comprehensive spending plan. Unexpectedly rapid progress in mapping prompted early revision of the original 5-year goals in the fall of 1993 [*Science* **262,** 43–46 (October 1, 1993)]. Current goals, which run through September 30, 1998, are listed on page 5; text of both 5-year plans is accessible via the Web (*http://www.ornl.gov/hgmis/project/hgp.html*).

DOE and NIH have adopted a joint policy to promote sharing of genome data and resources for facilitating progress and reducing duplicated work. (See Appendix B: DOE-NIH Sharing Guidelines, p. 75.)

ELSI Considerations

NIH and DOE devote at least 3% of their respective genome program budgets to identifying, analyzing, and addressing the ELSI considerations surrounding genome technology and the data it produces. The DOE ELSI component focuses on research into the privacy and confidentiality of personal genetic information, genetics relevant to the workplace, commercialization (including patenting) of genome research data, and genetic education for the general public and targeted communities. The NIH ELSI component supports studies on a range of ethical issues surrounding the conduct of genetic research and responsible clinical integration of new genetic technologies, especially in testing for mutations associated with cystic fibrosis and heritable breast, ovarian, and colon cancers.

In 1990, the DOE-NIH Joint ELSI Working Group was established to identify, address, and develop policy options; stimulate bioethics research; promote education of professional and lay groups; and collaborate with such international groups as the Human Genome Organisation (HUGO); United Nations Educational, Scientific, and Cultural Organization; and the European Community. Research funded by the U.S. Human Genome Project through the joint working group has produced policy recommendations in various areas. In May 1993, for example, the DOE-NIH Joint ELSI Working Group Task Force on Genetic Information and Insurance issued a report with recommendations for managing the impact of advances in human genetics on the current system of healthcare coverage. In 1996, the working group released guidelines for investigators on using DNA from human subjects for large-scale sequencing projects. The guidance emphasizes numerous ways to preserve donor anonymity [see Appendix C, p. 77, and the World Wide Web (*http://www.ornl.gov/hgmis/archive/nchgrdoe.html*)].

In 1997, following an evaluation, the two agencies modified the ELSI working group into the ELSI Research and Program Evaluation Group (ERPEG). ERPEG will focus more specifically on research activities supported by DOE and NIH ELSI programs.

Other U.S. Programs

The potential impact of genome research on society and the rapid growth of the biotechnology industry have spurred the initiation of other genome research projects in this country and worldwide. These projects aim to create maps of the human genome and the genomes of model organisms and several economically important microbes, plants, and animals.

- The DOE Microbial Genome Program, begun in 1994, is producing complete genome sequence data on industrially important microorganisms, including those that live under extreme environmental conditions. The sequences of several microbial genomes have been completed. *[http://www.er.doe.gov/production/ober/EPR/mig_top.html]*
- In 1990, the National Science Foundation, DOE, and the U.S. Department of Agriculture (USDA) initiated a project to map and sequence the genome of the model plant *Arabidopsis thaliana*. The goal of this project is to enhance fundamental understanding of plant processes. In 1996, the three agencies began funding systematic, large-scale genomic sequencing of the 120-megabase *Arabidopsis* genome, with the goal of completing it by 2004, with DOE support through the Office of Basic Energy Sciences. *[http://pgec-genome.pw.usda.gov/agi.html]*
- USDA also funds animal genome research projects designed to obtain genome maps for economically important species (e.g., corn, soybeans, poultry, cattle, swine, and sheep) to enable genetic modifications that will increase resistance to diseases and pests, improve nutrient value, and increase productivity.
- The Advanced Technology Program (ATP) of the U.S. National Institute of Standards and Technology promotes industry-government partnerships in DNA sequencing and biotechnology through the Tools for DNA Diagnostics component. DOE staff participates in the ATP review process (see box, p. 22). *[http://www.atp.nist.gov]*
- In 1997 the NIH National Cancer Institute established the Cancer Genome Anatomy Project (CGAP) to develop new diagnostic tools for understanding molecular changes that underlie all cancers (*http://www.ncbi.nlm.nih.gov/ncicgap*). DOE researchers are generating clone libraries to support this effort.

International Collaborations

The current DOE-NIH Five-Year Plan commends the "spirit of international cooperation and sharing" that has characterized the Human Genome Project and played a major role in its success. Cooperation includes collaborations among laboratories in the United States and abroad as well as extensive sharing of materials and information among genome researchers around the world. The DOE Human Genome Program supports many international collaborations as well as grantees in several foreign institutions.

Collaborations involving the DOE human genome centers include mapping chromosomes 16 and 19, developing resources, and constructing the human gene map from shared cDNA libraries. These libraries were generated by the Integrated Molecular Analysis of Gene Expression (called IMAGE) Consortium initiated by groups at Lawrence Livermore National Laboratory, Columbia University, NIH National Institute of Mental Health, and Généthon (France).

Investigators from almost every major sequencing center in the world met in Bermuda in February 1996 and again in 1997 to discuss issues related to large-scale sequencing. These meetings were designed to help researchers coordinate, compare, and evaluate human genome mapping and sequencing strategies; consider new sequencing and informatics technologies; and discuss release of data.

Human Genome Organisation

Founded by scientists in 1989, HUGO is a nongovernmental international organization providing coordination functions for worldwide genome efforts. HUGO activities range from support of data collation for constructing genome maps to organizing workshops. HUGO also fosters exchange of data and biomaterials, encourages technology sharing, and serves as a coordinating agency for building relationships among various government funding agencies and the genome community.

HUGO offers short-term (2- to 10-week) travel awards up to $1500 for investigators under age 40 to visit another country to learn new methods or techniques and to facilitate collaborative research between the laboratories.

HUGO has worked closely with international funding agencies to sponsor single-chromosome workshops (SCWs) and other genome meetings. Due to the success of these workshops as well as the shift in emphasis from mapping to sequencing, DOE and NIH began to phase out their funding for international SCWs in FY 1996 but encouraged applications for individual SCWs as needed. In 1996, HUGO partially funded an international strategy meeting in Bermuda on large-scale sequencing. Principles regarding data release and a resources list developed at the meeting are available on the HUGO Web site (*http://hugo.gdb.org/hugo.html*).

Membership in HUGO (over 1000 people in more than 50 countries) is extended to persons concerned with human genome research and related scientific subjects. Its current president is Grant R. Sutherland (Adelaide Women and Children's Hospital, Australia). Directed by an 18-member international council, HUGO is supported by grants from the Howard Hughes Medical Institute and The Wellcome Trust.

Countries with Genome Programs

Countries with genome programs or strong programs in human genetics include Australia, Brazil, Canada, China, Denmark, European Union, France, Germany, Israel, Italy, Japan, Korea, Mexico, Netherlands, Russia, Sweden, United Kingdom, and United States.

Los Alamos National Laboratory researchers Peter Goodwin and Rhett Affleck load a sample of fluorescently labeled DNA into an ultrasensitive flow cytometer used to detect single cleaved nucleotides. [Source: Lynn Clark, LANL]

Appendices

Appendix A: Early History, Enabling Legislation (1984–90) 73

Appendix B: DOE-NIH Sharing Guidelines (1992) 75

Appendix C: Human Subjects Guidelines (1996) 77

Appendix D: Genetics on the World Wide Web (1997) 83

Appendix E: 1996 Human Genome Research Projects (1996) 89

Appendix F: DOE BER Program (1997) 95

Appendix A
DOE Human Genome Program: Early History, Enabling Legislation

A brief history of the U.S. Department of Energy (DOE) Human Genome Program will be useful in a discussion of the objectives of the DOE program as well as those of the collaborative U.S. Human Genome Project. The DOE Office of Biological and Environmental Research (OBER) of DOE and its predecessor agencies—the Atomic Energy Commission and the Energy Research and Development Administration—have long sponsored research into genetics, both in microbial systems and in mammals, including basic studies on genome structure, replication, damage; and repair and the consequences of genetic mutations. (See Appendix E for a discussion of the DOE Biological and Environmental Research Program.)

In 1984, OBER [then named Office of Health and Environmental Research (OHER)] and the International Commission on Protection Against Environmental Mutagens and Carcinogens cosponsored a conference in Alta, Utah, which highlighted the growing roles of recombinant DNA technologies. Substantial portions of the meeting's proceedings were incorporated into the Congressional Office of Technology Assessment report, *Technologies for Detecting Heritable Mutations in Humans*, in which the value of a reference sequence of the human genome was recognized.

Acquisition of such a reference sequence was, however, far beyond the capabilities of biomedical research resources and infrastructure existing at that time. Although the small genomes of several microbes had been mapped or partially sequenced, the detailed mapping and eventual sequencing of 24 distinct human chromosomes (22 autosomes and the sex chromosomes X and Y) that together comprise an estimated 3 billion subunits was a task some thousandsfold larger.

DOE OHER was already engaged in several multidisciplinary projects contributing to the nation's biomedical capabilities, including the GenBank DNA sequence repository, which was initiated and sustained by DOE computer and data-management expertise. Several major user facilities supporting microstructure research were developed and are maintained by DOE. Unique chromosome-processing resources and capabilities were in place at Los Alamos National Laboratory and Lawrence Livermore National Laboratory. Among these were the fluorescence-activated cell sorter (called FACS) systems to purify human chromosomes within the National Laboratory Gene Library Project for the production of libraries of DNA clones. The availability of these monochromosomal libraries opened an important path—a practical means of subdividing the huge total genome into 24 much more manageable components.

With these capabilities, OHER began in 1986 to consider the feasibility of a dedicated human genome program. Leading scientists were invited to the March 1986 international conference at Santa Fe, New Mexico, to assess the desirability

Enabling Legislation

In the United States, the first federal support for genetics research was through the Atomic Energy Commission. In the early days of nuclear energy development, the focus was on radiation effects and later broadened under the Energy Research and Development Administration (ERDA) and the Department of Energy to include the health implications of all energy technologies and their by-products. Major enabling legislation follows.

Atomic Energy Act of 1946 (P.L. 79-585): Provided the initial charter for a comprehensive program of research and development related to the utilization of fissionable and radioactive materials for medical, biological, and health purposes.

Atomic Energy Act of 1954 (P.L. 83-703): Further authorized AEC "to conduct research on the biologic effects of ionizing radiation."

Energy Reorganization Act of 1974 (P.L. 93-438): Provided that responsibilities of ERDA should include "engaging in and supporting environmental, biomedical, physical, and safety research related to the development of energy resources and utilization technologies."

Federal Non-Nuclear Energy Research and Development Act of 1974 (P.L. 93-577): Authorized ERDA to conduct a comprehensive non-nuclear energy research, development, and demonstration program to include the environmental and social consequences of the various technologies.

DOE Organization Act of 1977 (P.L. 95-91): Instructed the department "to assure incorporation of national environmental protection goals in the formulation and implementation of energy programs; and to advance the goal of restoring, protecting, and enhancing environmental quality, and assuring public health and safety," and to conduct "a comprehensive program of research and development on the environmental effects of energy technology and programs."

and feasibility of implementing such a project. With virtual unanimity, participants agreed that ordering and eventually sequencing DNA clones representing the human genome were desirable and feasible goals. With the receipt of this enthusiastic response, OHER initiated several pilot projects. Program guidance was further sought from the DOE Health Effects Research Advisory Committee (HERAC).

HERAC Recommendation

The April 1987 HERAC report recommended that DOE and the nation commit to a large, multidisciplinary scientific and technological undertaking to map and sequence the human genome. DOE was particularly well suited to focus on resource and technology development, the report noted; HERAC further recommended a leadership role for DOE because of its demonstrated expertise in managing complex and long-term multidisciplinary projects involving both the development of new technologies and the coordination of efforts in industries, universities, and its own laboratories.

Evolution of the nation's Human Genome Project further benefited from a 1988 study by the National Research Council (NRC) entitled *Mapping and Sequencing the Human Genome*, which recommended that the United States support this research effort and presented an outline for a multiphase plan

DOE and NIH Coordination

The National Institutes of Health (NIH) was a necessary participant in the large-scale effort to map and sequence the human genome because of its long history of support for biomedical research and its vast community of scientists. This was confirmed by the NRC report, which recommended a major role for NIH. In 1987, under the leadership of Director James Wyngaarden, NIH established the Office of Genome Research in the Director's Office. In 1988, DOE and NIH signed a Memorandum of Understanding in which the agencies agreed to work together, coordinate technical research and activities, and share results. In 1990, DOE and NIH submitted a joint research plan outlining short- and long-term goals of the project.

Appendix B
DOE-NIH Guidelines for Sharing Data and Resources

*At its December 7, 1992, meeting, the DOE-NIH Joint Subcommittee on the Human Genome approved the following sharing guidelines, developed from the DOE draft of September 1991.**

The information and resources generated by the Human Genome Project have become substantial, and the interest in having access to them is widespread. It is therefore desirable to have a statement of philosophy concerning the sharing of these resources that can guide investigators who generate the resources as well as those who wish to use them.

A key issue for the Human Genome Project is how to promote and encourage the rapid sharing of materials and data that are produced, especially information that has not yet been published or may never be published in its entirety. Such sharing is essential for progress toward the goals of the program and to avoid unnecessary duplication. It is also desirable to make the fruits of genome research available to the scientific community as a whole as soon as possible to expedite research in other areas.

Although it is the policy of the Human Genome Project to maximize outreach to the scientific community, it is also necessary to give investigators time to verify the accuracy of their data and to gain some scientific advantage from the effort they have invested. Furthermore, in order to assure that novel ideas and inventions are rapidly developed to the benefit of the public, intellectual property protection may be needed for some of the data and materials.

After extensive discussion with the community of genome researchers, the advisors of the NIH and DOE genome programs have determined that consensus is developing around the concept that a 6-month period from the time the data or materials are generated to the time they are made available publicly is a reasonable maximum in almost all cases. More rapid sharing is encouraged.

Whenever possible, data should be deposited in public databases and materials in public repositories. Where appropriate repositories do not exist or are unable to accept the data or materials, investigators should accommodate requests to the extent possible.

The NIH and DOE genome programs have decided to require all applicants expecting to generate significant amounts of genome data or materials to describe in their application how and when they plan to make such data and materials available to the community. Grant solicitations will specify this requirement. These plans in each application will be reviewed in the course of peer review and by staff to assure they are reasonable and in conformity with program philosophy. If a grant is made, the applicant's sharing plans will become a condition of the award and compliance will be reviewed before continuation funding is provided. Progress reports will be asked to address the issue.

---

*Reprinted from *Human Genome News* **4**(5), 4 (1993).

Appendix C
NIH-DOE Guidance on Human Subjects Issues in Large-Scale DNA Sequencing

Date issued: August 9, 1996

Introduction

The Human Genome Project (HGP) is now entering into large-scale DNA sequencing. To meet its complete sequencing goal, it will be necessary to recruit volunteers willing to contribute their DNA for this purpose. The guidance provided in this document is intended to address ethical issues that must be considered in designing strategies for recruitment and protection of DNA donors for large-scale sequencing.

Nothing in this document should be construed to differ from, or substitute for, the policies described in the Federal Regulations for the Protection of Human Subjects [45CFR46 (NIH) and 10CFR745 (DOE)]. Rather, it is intended to supplement those policies by focusing on the particular issues raised by large-scale human DNA sequencing. This statement addresses six topics: (1) benefits and risks of genomic DNA sequencing; (2) privacy and confidentiality; (3) recruitment of DNA donors as sources for library construction; (4) informed consent; (5) IRB approval; and (6) use of existing libraries.

The guidance provided in this statement is intended to afford maximum protection to DNA donors and is based on the belief that protection can best be achieved by a combination of approaches including:

- ensuring that the initial version of the complete human DNA sequence is derived from multiple donors;
- providing donors with the opportunity to make an informed decision about whether to contribute their DNA to this project; and
- taking effective steps to ensure the privacy and confidentiality of donors.

1. Benefits and Risks of Genomic DNA Sequencing

The HGP offers great promise for the improvement of human health. As a consequence of the HGP, there will be a more thorough understanding of the genetic bases of human biology and of many diseases. This, in turn, will lead to better therapies and, perhaps more importantly, prevention strategies for many of those diseases. Similarly, as the technology developed by the HGP is applied to understanding the biology of other organisms, many other human activities will be affected including agriculture, environmental management, and biologically based industrial processes.

While the HGP offers great promise to humanity, there will be no direct benefit, in either clinical or financial terms, to any of the individuals who choose to donate DNA for large-scale sequencing. Rather, the motivation for donation is likely to be an altruistic willingness to contribute to this historic research effort.

However, individuals who donate DNA to this effort may face certain risks. Information derived from the donors will become available in public databases. Such information may reveal, for example, DNA sequence-based information about disease susceptibility. If the donor becomes aware of such information, it could lead to emotional distress on her/his part. If such health-related information becomes known to others, discrimination against the donor (e.g., in insurance or in employment) could result. Unwanted notoriety is another potential risk to donors. Therefore, those engaged in large-scale sequencing must be sensitive to the unique features of this type of research and ensure that both the protections normally afforded research subjects and the special issues associated with human genomic DNA sequencing are thoroughly addressed.

While some risks to donors can already be identified, the probability of adverse events materializing appears to be low. However, the risks of harm to individuals will increase if confidentiality is not maintained and/or the number of donors is limited to a very few individuals. Either, or both, of these situations would increase the possibility of a donor's identity being revealed without his/her knowledge or permission.

A final issue to consider is characterized in a statement taken from the OPRR Guidebook[1] which points out that "some areas [of genetic research] present issues for which no clear guidance can be given at this point, either because enough is not known about the risks presented by the research, or because no consensus on the appropriate resolution of the problem exists." It is anticipated that the DNA sequence information produced by the Human Genome Project will be used in the future for types of research which cannot now be predicted and the risks of which cannot be assessed or disclosed.

2. Privacy and Confidentiality

In general, one of the most effective ways of protecting volunteers from the unexpected, unwelcome or unauthorized use of information about them is to ensure that there are no opportunities for linking an individual donor with information about him/her that is revealed by the research. By not collecting information about the identity of a research subject and any biological material or records developed in the course of the research, or by subsequently removing all identifiers ("anonymizing" the sample), the possibility of risk to the subject stemming from the results of the research is greatly reduced. Large-scale DNA sequence determination represents an exception because each person's DNA sequence is unique and, ultimately, there is enough information in any individual's DNA sequence to absolutely identify her/him. However, the technology that would allow the unambiguous identification of an individual from his/her DNA sequence is not yet mature. Thus, for the foreseeable future, establishing effective confidentiality, rather than relying on anonymity, will be a very useful approach to protecting donors.

Investigators should introduce as many disconnects between the identity of donors and the publicly available information and materials as possible. There should not be any way for anyone to establish that a specific DNA sequence came from a particular individual, other than resampling an individual's DNA and comparing it to the sequence information in the public database. In particular, no phenotypic or demographic information about donors should be linked to the DNA to be sequenced.[2] For the purposes of the HGP such information will rarely be useful, and recording such information could result in possible misuse and compromise donor confidentiality.

Confidentiality should be "two way." Not only should others be unable to link a DNA sequence to a particular individual, but no individual who donates DNA should be able to confirm directly that a particular DNA sequence was obtained from their DNA sample.[3] This degree of confidentiality will preclude the possibility of re-contacting DNA donors, providing another degree of protection for them. It should be clear to both investigators and to donors that the contact involved in obtaining the initial specimen will be the only contact.[4]

Another approach for protecting all DNA donors is to reduce the incentive for wanting to know the identities of particular donors. If the initial human sequence is a "mosaic" or "patchwork" of sequenced regions derived from a number of different individuals, rather than that of a single individual, there would be considerably less interest in who the specific donors were. Although there may be scientific justification that each clone library used for sequencing should be derived from one person, there is no scientific reason that the entire initial human DNA sequence should be that of a single individual. As approximately 99.9% of the human DNA sequence is common between any two individuals, most of the fundamental biological information contained in the human DNA sequence is common to all people.

To increase the likelihood that the first human DNA sequence will be an amalgam of regions sequenced from different sources, a number of clone libraries must be made available. Although a number of large insert libraries have been made, most do not meet all of the standards set in this document; therefore, these libraries should be used as substrates for large-scale sequencing only under circumscribed conditions (see section 6, p. 79). Starting immediately, new libraries will be developed that have the advantage of being constructed in accordance with the ethical principles discussed in this document; they may also confer some additional scientific benefit. Such libraries are critical for the long-range needs of the HGP.

3. Source/Recruitment of DNA Donors for Library Construction

Another implication of the fact that 99.9% of the human DNA sequence is shared by any two individuals is that the backgrounds of the individuals who donate DNA for the first human sequence will make no scientific difference in terms of the usefulness and applicability of the information that results from sequencing the human genome. At the same time, there will undoubtedly be some sensitivity about the choice of DNA sources. There are no scientific reasons why DNA donors should not be selected from diverse pools of potential donors.[5]

There are two additional issues that have arisen in considering donor selection. These warrant particular discussion:

- It is recognized that women have historically been underrepresented in research, so it can be anticipated that concerns might arise if males (sperm DNA) were used exclusively as the source of DNA for large-scale sequencing. Although there would be no scientific basis for concern, because even in the case of a male source, half of the donor's DNA would have come from his mother and half from his father, nevertheless perceptions are not to be dismissed. While the choice of donors will not be dictated to investigators, it is expected that, because multiple libraries will be produced, a number of them will be made from female sources while others will be made from male sources.
- Staff of laboratories involved in library construction and DNA sequencing may be eager to volunteer to be donors because of their interest and belief in the HGP. However, proximity to the research may create some special vulnerabilities for laboratory staff members. It is also possible that they will feel pressure to donate and there may be an increased likelihood that confidentiality would be breached. Finally, there is a potential that the choice of persons so closely involved in the research may be interpreted as elitist. For all of these reasons, it is recommended that donors should not be recruited from laboratory staff, including the principal investigator.

4. Informed Consent

Obtaining informed consent specifically for the purpose of donating DNA for large-scale sequencing raises some unique concerns. Because anonymity cannot be guaranteed and confidentiality protections are not absolute, the disclosure process to potential donors must clearly specify what the process of DNA donation involves, what may make it different from other types of research, and what the implications are of one's DNA sequence information being a public scientific resource.

Federal regulations (45CFR46 and 10CFR745) require the disclosure of a number of issues in any informed consent document. They include such issues as potential benefits of the research, potential risks to the donor, control and ownership of donated material, long-term retention of donated material for future use, and the procedures that will be followed. In addition, there are several other disclosures that are of special importance for donors of DNA for large-scale sequencing. These include:

- the meaning of confidentiality and privacy of information in the context of large-scale DNA sequencing, and how these issues will be addressed;
- the lack of opportunity for the donor to later withdraw the libraries made from his/her DNA or his/her DNA sequence information from public use;
- the absence of opportunity for information of clinical relevance to be provided to the donor or her/his family;
- the possibility of unforeseen risks; and
- the possible extension of risk to family members of the donor or to any group or community of interest (e.g., gender, race, ethnicity) to which a donor might belong.

Many academic human genetics units have considerable experience in dealing with research subjects and obtaining informed consent, while the laboratories that are likely to be involved in making the libraries for sequencing have, in general, much less experience of this type. Therefore, library makers are encouraged to establish a collaboration with one or more human genetics units, with the latter being responsible for recruiting donors, obtaining informed consent, obtaining the necessary biological samples, and providing a blinded sample to the library maker. Collaboration with tissue banks may be considered as long as these banks are collecting tissues in accordance with this guidance. The library maker should have no contact with the donor and no opportunity to obtain any information about the donor's identity.

5. IRB Approval

Effective immediately, projects to construct libraries for large-scale DNA sequencing must obtain Institutional Review Board (IRB) approval before work is initiated. IRBs should carefully consider the unique aspects of large-scale sequencing projects. Some of the informed consent provisions outlined may be somewhat at odds with the usual and customary disclosures found in most protocols involving human subjects and which IRBs usually consider. For example, research subjects usually are given the opportunity to withdraw from a research project if they change their minds about participating. In the case of donors for large-scale sequencing, it will not be possible to withdraw either the libraries made from their DNA or the DNA sequence information obtained using those libraries once the information is in the public domain. By the time a significant amount of DNA sequence data has been collected, the libraries, as well as individual clones from them, will have been widely distributed and the sequence information will have been deposited in and distributed from public databases. In addition, there will be no possibility of returning information of clinical relevance to the donor or his/her family.

6. Use of Existing Libraries for Large-Scale Sequencing

Many of the existing libraries (including those derived from anonymous donors) were not made in complete conformity with the principles elaborated above. The potential risks that may result from their use will be minimized by the rapid introduction of several new libraries constructed in accordance with this guidance, which NCHGR and DOE are taking steps to initiate. This will ensure that the existing libraries will only contribute small amounts to the first complete human DNA sequence. In the interim, existing libraries can continue to be used for large-scale sequencing, only if IRB approval and consent for "continued use" are obtained[6] and approval by the funding agency is granted.

It is important that in obtaining consent for contined use of existing libraries, no coercion of the DNA donor occur. It is therefore recommended that consideration be given to whether it is appropriate for the individual who previously recruited the donor to recontact him/her to obtain this consent. In some cases an IRB may determine that the recontact should be made by a third party to assure that the donors are fully informed and allowed to choose freely whether their DNA can continue to be used for this purpose.

Conclusion

This document is intended to provide guidance to investigators and IRBs who are involved in large-scale sequencing efforts. It is designed to alert them to special ethical concerns that may arise in such projects. In particular, it provides guidance for the use of existing and the construction of new DNA libraries. Adhering to this guidance will ensure that the initial version of the complete human sequence is derived from multiple, diverse donors; that donors will have the opportunity to make an informed decision about whether to contribute their DNA to this project; and that effective steps will be taken by investigators to ensure the privacy and confidentiality of donors.

Investigators funded by NCHGR and DOE to develop new libraries for large-scale human DNA sequencing will be required to have their plans for the recruitment of DNA donors, including the informed consent documents, reviewed and approved by the funding agency before donors are recruited. Investigators involved in large-scale human sequencing will also be asked to observe those aspects of this guidance that pertain to them.

Approved August 17, 1996, by:

Francis S. Collins, M.D., Ph.D., Director, National Center for Human Genome Research, National Institutes of Health
Aristides N. Patrinos, Ph.D., Associate Director, Office of Health and Environmental Research, U.S. Department of Energy

---

Footnotes

1. Office of Protection from Research Risks, Protecting Human Research Subjects: Institutional Review Board Guidebook (OPRR: U.S. Government Printing Office, 1993).

2. It is recognized that it will be trivially easy to determine the sex of the donor of the library, by assaying for the presence or absence of Y chromosome in the library.

3. There are a number of approaches to preventing a DNA donor from knowing that his/her DNA was actually sequenced as part of the HGP. For example, each time a clone library is to be made, an appropriately diverse pool of between five and ten volunteers can be chosen in such a way that none of them knows the identity of anyone else in the pool. Samples for DNA preparation and for preparation of a cell line can be collected from all of the volunteers (who have been told that their specimen may or may not eventually be used for DNA sequencing) and one of those samples is randomly and blindly selected as the source actually used for library construction. In this way, not only will the identity of the individual whose DNA is chosen not be known to the investigators, but that individual will also not be sure that s/he is the actual source.

4. Although recontacting donors should not be possible, investigators will potentially want to be able to resample a donor's genome. Thus, at the time the initial specimen is obtained, in addition to making a clone library representing the donor's genome, it should also be used to prepare an additional aliquot of high molecular weight DNA for storage and a permanent cell line. Either resource could then be used as source of the donor's genome in case additional DNA were needed or comparison with the results of the analysis of the cloned DNA were desired.

5. There has been discussion in the scientific communit about the sex of DNA donors. A library prepared from a female donor will contain DNA from the X chromosome in an amount equivalent to the autosomes, but will completely lac Y chromosomal DNA. Conversely, a library prepared from a male donor will contain Y DNA, but both X and Y DNA wil only be present at half the frequency of the DNA from the other chromosomes. Scientifically, then, there are both advantages and disadvantages inherent in the use of either a male or a female donor. The question of the sex of the donor also involves the question of the use of somatic or germ line DNA to make libraries. For making libraries, useful amounts of germ line DNA can only be obtained from a male source (i.e., from sperm); it is not possible to obtain enough ova from a female donor to isolate germ line DNA for this purpose. Opinion is divided in the scientific community about whether germ line or somatic DNA should be used for large-scale sequencing. Somatic DNA is known to be rearranged, relative to germ line DNA, in certain regions (e.g., the immunoglobulin genes) and the possibility has been raised that other developmentally based rearrangements may occur, although no example of the latter has been offered. While some believe that the sequence product should not contain any rearrangements of this sort, others consider this potential advantage of germ line DNA to be relatively minor in comparison to the need to have the X chromosome fully represented in sequencing efforts and prefer the use of somatic DNA.

6. Individuals whose DNA was used for library construction (with the exception of those created from deceased or anonymous individuals) should be fully informed about the risks and benefits described above, should freely choose whether they would like their DNA to continue to be used fo this purpose, and their decision should be documented.

Executive Summary of Joint NIH-DOE Human Subjects Guidance

1. Those engaged in large-scale sequencing must be sensitive to the unique features of this type of research and ensure that both the protections normally afforded research subjects and the special issues associated with human genomic DNA sequencing are thoroughly addressed.

2. For the foreseeable future, establishing effective confidentiality, rather than relying on anonymity, will be a very useful approach to protecting donors.

3. Investigators should introduce as many disconnects between the identity of donors and the publicly available information and materials as possible.

4. No phenotypic or demographic information about donors should be linked to the DNA to be sequenced.

5. There are no scientific reasons why DNA donors should not be selected from diverse pools of potential donors.

6. While the choice of donors will not be dictated to investigators, it is expected that, because multiple libraries will be produced, a number of them will be made from female sources while others will be made from male sources.

7. It is recommended that donors should not be recruited from laboratory staff, including the principal investigator.

8. The disclosure process to potential donors must clearly specify what the process of DNA donation involves, what may make it different from other types of research, and what the implications are of one's DNA sequence information being a public scientific resource.

9. Library makers are encouraged to establish a collaboration with one or more human genetics units [or tissue banks].

10. The library maker should have no contact with the donor and no opportunity to obtain any information about the donor's identity.

11. Effective immediately, projects to construct libraries for large-scale DNA sequencing must obtain Institutional Review Board (IRB) approval before work is initiated.

12. Existing libraries can continue to be used for large-scale sequencing, only if IRB approval and consent for continued use are obtained and approval by the funding agency is granted.

13. It is important that in obtaining informed consent for continued use of existing libraries, no coercion of the DNA donor occur.

Appendix D
Human Genome Project and Genetics on the World Wide Web

August 1997

The World Wide Web offers the easiest path to information about the Human Genome Project and related genetics topics. Some useful sites to visit are included in the list below.

Human Genome Project

DOE Human Genome Program
http://www.er.doe.gov/production/ober/hug_top.html

Devoted to the DOE component of the U.S. Human Genome Project and to the DOE Microbial Genome Program. Links to many other sites.

Human Genome Project Information
http://www.ornl.gov/hgmis

Comprehensive site covering topics related to the U.S. and worldwide Human Genome Projects. Useful for updating scientists and providing educational material for nonscientists, in support of DOE's commitment to public education. Developed and maintained for DOE by the Human Genome Management Information System (HGMIS) at Oak Ridge National Laboratory.

NIH National Human Genome Research Institute
http://www.nhgri.nih.gov

Site of the NIH sector of the U.S. Human Genome Project.

DOE Human Genome Program Publications

***Human Genome News**
http://www.ornl.gov/hgmis/publicat/publications.html

Quarterly newsletter reporting on the worldwide Human Genome Project.

Biological Sciences Curriculum Study (BSCS) Teaching Modules

Online versions in preparation; hardcopies available from 719/531-5550

- "Genes, Environment, and Human Behavior," tentative title, in preparation
- "Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy" (1992)
- "The Human Genome Project: Biology, Computers, and Privacy" (1996)
- "The Puzzle of Inheritance: Genetics and the Methods of Science" (1997)

***Primer on Molecular Genetics, 1992**
http://www.ornl.gov/hgmis/publicat/publications.html#primer

Explains the science behind the genome research.

***To Know Ourselves, 1996**
http://www.ornl.gov/hgmis/tko

Booklet reviewing DOE's role, history, and achievements in the Human Genome Project and introducing the science and other aspects of the project.

Ethical, Legal, and Social Issues Related to Genetics Research

HGMIS Gateways Web page
http://www.ornl.gov/hgmis/links.html

Choose "Ethical, Legal, and Social Issues."

Center for Bioethics, University of Pennsylvania
http://www.med.upenn.edu/~bioethic

Full-text articles about such ethical issues as human cloning; includes a primer on bioethics.

Courts and Science On-Line Magazine (CASOLM)
http://www.ornl.gov/courts

Coverage of genetic issues affecting the courts.

ELSI in Science
http://www.lbl.gov/Education/ELSI/ELSI.html

Teaching modules designed to stimulate discussion on implications of scientific research.

Eubios Ethics Institute
http://www.biol.tsukuba.ac.jp/~macer/index.html

Site includes newsletter summarizing literature in bioethics and biotechnology.

Genetic Privacy Act
http://www.ornl.gov/hgmis/resource/elsi.html

Model legislation written with support of the DOE Human Genome Program.

MCET—The Human Genome Project
http://phoenix.mcet.edu/humangenome/index.html

ELSI issues for high school students.

---

*Print copy available from HGMIS (see p. 87 or inside front cover for contact information).

National Bioethics Advisory Committee
http://www.nih.gov/nbac/nbac.htm

The bioethics committee offers advice to the National Science and Technology Council and others on bioethical issues arising from research related to human biology and behavior.

National Center for Genomic Resources
http://www.ncgr.org

Comprehensive Genetics and Public Issues page; includes congressional bills related to genetic privacy.

The Gene Letter
http://www.geneletter.org/genetalk.html

Bimonthly newsletter to inform consumers and professionals about advances in genetics and encourage discussion about emerging policy dilemmas.

Your Genes, Your Choices
http://www.nextwave.org/ehr/books/index.html

Booklet written in simple English, describing the Human Genome Project; the science behind it; and how ethical, legal, and social issues raised by the project may affect people's everyday lives.

General Genetics and Biotechnology

Many of the following sites contain links to both educational and technical material.

HGMIS Community Education and Outreach Gateways Web Page
http://www.ornl.gov/hgmis/links.html

Access Excellence
http://outcast.gene.com/ae/index.html

Extensive genetic and biotechnology resources for teachers and nonscientists.

BIO Online (Biotechnology Industry Organization)
http://www.bio.com

Comprehensive directory of biotechnology sites on the Internet.

Biospace
http://www.biospace.com

Biotech industry site; profiles biotech companies by region.

BioTech
http://biotech.chem.indiana.edu

An interactive educational resource and biotech reference tool; includes a dictionary of 6000 life science terms.

Biotechnology Information Center, USDA National Agricultural Library
http://www.nal.usda.gov/bic

Comprehensive agricultural biotechnology resource; includes a bibliography on patenting biotechnology products and processes (*http://www.nal.usda.gov/bic/Biblios/patentag.htm*).

Bugs 'N Stuff
http://www.ncgr.org/microbe

List of microbial genomes being sequenced, research groups, genome sizes, and facts about selected organisms. Links to related sites.

Careers in Genetics
http://www.faseb.org/genetics/gsa/careers/bro-menu.htm

Online booklet from the Genetics Society of America, including several profiles of geneticists. See also career sections of sites specified above, such as Access Excellence.

Carolina Biological Supply Company
http://www.carosci.com/Tips.htm

Teaching materials for all levels. Includes mini-lessons on selected scientific topics, two online magazines, What's New, software, catalogs, and publications.

Cell & Molecular Biology Online
http://www.tiac.net/users/pmgannon

Links to electronic publications, current research, educational and career resources, and more.

CERN Virtual Library, Genetics section, Biosciences Division
http://www.ornl.gov/TechResources/Human_Genome/genetics.html

Includes an organism index linking to other pertinent databases, information on the U.S. and international Human Genome Projects, and links to research sites.

Classic Papers in Genetics
http://www.esp.org

Covers the early years, with introductory notes. See also Access Excellence site above for genetics history.

Community of Science Web Server
http://cos.gdb.org/best.html

Links to Medline, U.S. Patent Citation Database, Commerce Business Daily, The Federal Register, and other resources.

Database of Genome Sizes
http://www.cbs.dtu.dk/databases/DOGS/index.html

Lists numerous organisms with genome sizes, scientific and common names, classifications, and references.

Genetic and biological resources links
http://www.er.doe.gov/production/ober/bioinfo_center.html

Genetics Education Center, University of Kansas Medical Center
http://www.kumc.edu/instruction/medicine/genetics/homepage.html

Educational information on human genetics, career resources.

Genetics Glossary
http://www.ornl.gov/hgmis/publicat/glossary.html

Glossary of terms related to genetics.

Genetics Webliography
http://www.dml.georgetown.edu/%7Edavidsol/len.html

Extensive links for researchers and nonscientists from Georgetown University Library.

Genomics: A Global Resource
http://www.phrma.org/genomics/index.html

Many links. Website a joint project of the Pharmaceutical Research and Manufacturers of America and the American Institute of Biological Sciences; includes Genomics Today, a daily update on the latest news in the field.

Hispanic Educational Genome Project
http://vflylab.calstatela.edu/hgp

Designed to educate high school students and their families about genetics and the Human Genome Project. Links to other projects.

Howard Hughes Medical Institute
http://www.hhmi.org

Home page of major U.S. philanthropic organization that supports research in genetics, cell biology, immunology, structural biology, and neuroscience. Excellent introductory information on these topics.

Library of Congress
http://lcweb.loc.gov/homepage/lchp.html

Microbial Database
http://www.tigr.org/tdb/mdb/mdb.html

Lists completed and in-progress microbial genomes, with funding sources.

MIT Biology Hypertextbook
http://esg-www.mit.edu:8001/esgbio/7001main.html

All the basics.

Science and Mathematics Resources
http://www-sci.lib.uci.edu

More than 2000 Web references, including Frank Potter's Science Gems and Martindale's Health Science Guide. For teachers at all levels.

Virtual Courses on the Web
http://lenti.med.umn.edu/~mwd/courses.html

Links to Web tutorials in biology, genetics, and more.

Welch Web
http://www.welch.jhu.edu

Links to many Internet biomedical resources, dictionaries, encyclopedias, government sites, libraries, and more, from the Johns Hopkins University Welch Library.

Why Files
http://whyfiles.news.wisc.edu

Illustrated explanations of the science behind the news.

Images on the Web

Biochemistry Online
http://biochem.arach-net.com

Essays, courses, 3-D images of biomolecules, modeling, software.

Bugs in the News!
http://falcon.cc.ukans.edu/~jbrown/bugs.html

Microbiology information and a nice collection of images of biological molecules.

Cells Alive!
http://www.cellsalive.com

Images (some moving) of different types of cells.

Cn3D (See in 3-D)
http://www3.ncbi.nlm.nih.gov/Entrez/Structure/cn3d.html

3-D molecular structure viewer allowing the user to visualize and rotate structure data entries from Entrez. Highly technical, for researchers.

Cytogenetics Gallery
http://www.pathology.washington.edu:80/Cytogallery

Photos (karyotypes) of normal and abnormal chromosomes.

DNA Learning Center, Cold Spring Harbor Laboratory
http://darwin.cshl.org/index.html

Animated images of PCR and Southern Blotting techniques.

Gene Map from the 1996 Genome Issue of Science
http://www.ncbi.nlm.nih.gov/SCIENCE96

Click on particular areas of chromosomes and find genes.

Images of Biological Molecules
http://www.cc.ukans.edu/~micro/picts.html

3-D structures of proteins and nucleic acids obtained from Brookhaven National Laboratory Protein Database and others.

Lawrence Livermore National Laboratory Chromosome 19 Physical Map
http://www-bio.llnl.gov/bbrp/genome/genome.html

Los Alamos National Laboratory Chromosome 16 Physical Map
http://www-ls.lanl.gov/DBqueries/QueryPage.html

Journals and Magazines

HGMIS Journals Gateways Web page
http://www.ornl.gov/hgmis/links.html

Choose "Journals, Books, Periodicals."

Biochemistry and Molecular Biology Journals
http://biochem.arach-net.com/beasley/journals.html

Comprehensive list.

Nature, Nature Genetics, and Nature Biotechnology
http://www.nature.com

Abstracts of articles, full text of letters and editorials.

Science Magazine
http://www.sciencemag.org

Abstracts and some full-text articles.

Science Magazine Genome Issue (10/96)
http://www.sciencemag.org/science/content/vol274/issue5287

Full text includes a "clickable" gene map.

Science News
http://www.sciencenews.org

Online version of weekly popular science magazine with full text of selected articles.

Medical Genetics

Blazing a Genetic Trail
http://www.hhmi.org/GeneticTrail

Illustrated booklet from the Howard Hughes Medical Institute on hunting for disease genes.

Directory of National Genetic Voluntary Organizations and Related Resources
http://medhlp.netusa.net/agsg/agsgsup.htm

Support groups for people with genetic diseases and their families.

GeneCards
http://bioinformatics.weizmann.ac.il/cards

A database of more than 6000 genes; describes their functions, products, and biomedical applications.

Gene Therapy
http://www.mc.vanderbilt.edu/gcrc/gene/index.html

Web course covering the basics, with links to other sites.

Inherited-Disease Genes Found by Positional Cloning
http://www.ncbi.nlm.nih.gov/Baxevani/CLONE/index.html

Links to OMIM.

NIH Office of Recombinant DNA Activities
http://www.nih.gov/od/orda

Includes a database of human gene therapy protocols.

Online Mendelian Inheritance in Man (OMIM)
http://www.ncbi.nlm.nih.gov/Omim

A comprehensive, authoritative, and up-to-date human gene and genetic disorder catalog that supports medical genetics and the Human Genome Project.

Promoting Safe and Effective Genetic Testing in the United States (1997)
http://www.med.jhu.edu/tfgtelsi

Principles and recommendations by a joint NIH-DOE Human Genome Project group that examined the development and provision of gene tests in the United States.

Understanding Gene Testing
http://www.gene.com/ae/AE/AEPC/NIH/index.html

Illustrated brochure from the National Cancer Institute.

Science in the News

EurekAlert! *http://www.eurekalert.org*

InScight: *http://www.apnet.com/inscight*

SciWeb: *http://www.sciweb.com/news.html*

Short summaries of major stories, some with links to related articles in other sources.

HMS Beagle
http://biomednet.com/hmsbeagle

Biweekly electronic journal featuring major science stories, profiles, book reviews, and other items of interest.

Science Daily
http://www.sciencedaily.com

Headline stories, articles, and links to news services, newspapers, magazines, broadcast sources, journals, and organizations. Also offers weekly bulletins for updates by e-mail.

Science Guide
http://www.scienceguide.com

Daily news and information service and free science news e-mailer. Also contains directories of newsgroups, grant and funding resources, employment, and online journals.

ScienceNow
http://www.sciencenow.org

Daily online news service from Science magazine offers articles on major science news.

Web Search Tools

Biosciences Index to WWW Virtual Library
http://golgi.harvard.edu/htbin/biopages

Metacrawler
http://www.metacrawler.com

"Search the Net"
http://metro.turnpike.net/adorn/search.html

Comprehensive list of search tools, libraries, world fact books, and other useful information.

Search.com
http://www.search.com

Yahoo!
http://www.yahoo.com

Prepared August 1997 by
Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
423/576-6669, *caseydk@ornl.gov*
http://www.ornl.gov/hgmis

Appendix E
1996 Human Genome Research Projects

Research abstracts of these projects appear in Part 2 of this report.

Sequencing

Advanced Detectors for Mass Spectrometry
W.H. Benner and **J.M. Jaklevic**
Lawrence Berkeley National Laboratory, Berkeley, California

Mass Spectrometer for Human Genome Sequencing
Chung-Hsuan Chen
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Genomic Sequence Comparisons
George Church
Harvard Medical School, Boston, Massachusetts

A PAC/BAC End-Sequence Data Resource for Sequencing the Human Genome: A 2-Year Pilot Study
Pieter de Jong
Roswell Park Cancer Institute, Buffalo, New York

Multiple-Column Capillary Gel Electrophoresis
Norman Dovichi
University of Alberta, Edmonton, Canada

DNA Sequencing with Primer Libraries
John J. Dunn and **F. William Studier**
Brookhaven National Laboratory, Upton, New York

Rapid Preparation of DNA for Automated Sequencing
John J. Dunn and **F. William Studier**
Brookhaven National Laboratory, Upton, New York

A PAC/BAC End-Sequence Database for Human Genomic Sequencing
Glen A. Evans
University of Texas Southwestern Medical Center, Dallas, Texas

Automated DNA Sequencing by Parallel Primer Walking
Glen A. Evans
University of Texas Southwestern Medical Center, Dallas, Texas

*Parallel Triplex Formation as Possible Approach for Suppression of DNA-Viruses Reproduction
V.L. Florentiev
Russian Academy of Sciences, Moscow, Russia

Advanced Automated Sequencing Technology: Fluorescent Detection for Multiplex DNA Sequencing
Raymond F. Gesteland
University of Utah, Salt Lake City, Utah

Resource for Molecular Cytogenetics
Joe Gray and **Daniel Pinkel**
University of California, San Francisco

DNA Sample Manipulation and Automation
Trevor Hawkins
Whitehead Institute and Massachusetts Institute of Technology, Cambridge, Massachusetts

Construction of a Genome-Wide Characterized Clone Resource for Genome Sequencing
Leroy Hood, Mark D. Adams,[1] and **Melvin Simon**[2]
University of Washington, Seattle
[1]The Institute for Genomic Research, Rockville, Maryland
[2]California Institute of Technology, Pasadena, California

DNA Sequencing Using Capillary Electrophoresis
Barry L. Karger
Northeastern University, Boston, Massachusetts

Ultrasensitive Fluorescence Detection of DNA
Richard A. Mathies and **Alexander N. Glazer**
University of California, Berkeley

Joint Human Genome Program Between Argonne National Laboratory and the Engelhardt Institute of Molecular Biology
Andrei Mirzabekov
Argonne National Laboratory, Argonne, Illinois, and
Engelhardt Institute of Molecular Biology, Moscow, Russia

High-Throughput DNA Sequencing: SAmple SEquencing (SASE) Analysis as a Framework for Identifying Genes and Complete Large-Scale Genomic Sequencing
Robert K. Moyzis
Los Alamos National Laboratory, Los Alamos, New Mexico

One-Step PCR Sequencing
Barbara Ramsay Shaw
Duke University, Durham, North Carolina

---

*Projects designated by an asterisk were funded through small emergency grants to Russian scientists following December 1992 site reviews by David Galas (formerly of OHER, renamed OBER in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (LLNL).

Automation of the Front End of DNA Sequencing

Lloyd M. Smith and **Richard A. Guilfoyle**
University of Wisconsin, Madison

High-Speed DNA Sequence Analysis by Matrix-Assisted Laser Desorption Mass Spectrometry

Lloyd M. Smith
University of Wisconsin, Madison

Analysis of Oligonucleotide Mixtures by Electrospray Ionization-Mass Spectrometry

Richard D. Smith
Pacific Northwest National Laboratory, Richland, Washington

High-Speed Sequencing of Single DNA Molecules in the Gas Phase by FTICR-MS

Richard D. Smith
Pacific Northwest National Laboratory, Richland, Washington

Characterization and Modification of DNA Polymerases for Use in DNA Sequencing

Stanley Tabor
Harvard University, Boston, Massachusetts

Modular Primers for DNA Sequencing

Levy Ulanovsky[1,2]
[1]Argonne National Laboratory, Argonne, Illinois
[2]Weizmann Institute of Science, Rehovot, Israel

Time-of-Flight Mass Spectroscopy of DNA for Rapid Sequence

Peter Williams
Arizona State University, Tempe, Arizona

Development of Instrumentation for DNA Sequencing at a Rate of 40 Million Bases Per Day

Edward S. Yeung
Iowa State University, Ames, Iowa

Mapping

Resolving Proteins Bound to Individual DNA Molecules

David Allison and **Bruce Warmack**
Oak Ridge National Laboratory, Oak Ridge, Tennessee

*Improved Cell Electrotransformation by Macromolecules

Alexandre S. Boitsov
St. Petersburg State Technical University, St. Petersburg, Russia

Overcoming Genome Mapping Bottlenecks

Charles R. Cantor
Boston University, Boston, Massachusetts

Preparation of PAC Libraries

Pieter J. de Jong
Roswell Park Cancer Institute, Buffalo, New York

Chromosomes by Third-Strand Binding

Jacques R. Fresco
Princeton University, Princeton, New Jersey

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao
Eleanor Roosevelt Institute for Cancer Research, Denver, Colorado

*Identification and Mapping of DNA-Binding Proteins Along Genomic DNA by DNA-Protein Crosslinking

V.L. Karpov
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

A PAC/BAC Data Resource for Sequencing Complex Regions of the Human Genome: A 2-Year Pilot Study

Julie R. Korenberg
Cedars Sinai Medical Center, Los Angeles, California

Mapping and Sequencing of the Human X Chromosome

D. L. Nelson
Baylor College of Medicine, Houston, Texas

*Sequence-Specific Proteins Binding to the Repetitive Sequences of High Eukaryotic Genome

Olga Podgornaya
Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia

*Protein-Binding DNA Sequences

O.L. Polanovsky
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

*Development of Intracellular Flow Karyotype Analysis

A.I. Poletaev
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

Mapping and Sequencing with BACs and Fosmids

Melvin I. Simon
California Institute of Technology, Pasadena, California

Towards a Globally Integrated, Sequence-Ready BAC Map of the Human Genome

Melvin I. Simon
California Institute of Technology, Pasadena, California

Generation of Normalized and Subtracted cDNA Libraries to Facilitate Gene Discovery

Marcelo Bento Soares
Columbia University, New York, New York

Mapping in Man-Mouse Homology Regions

Lisa Stubbs
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Positional Cloning of Murine Genes

Lisa Stubbs
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Human Artificial Episomal Chromosomes (HAECS) for Building Large Genomic Libraries

Jean-Michel H. Vos
University of North Carolina, Chapel Hill

*Cosmid and cDNA Map of a Human Chromosome 13q14 Region Frequently Lost at B Cell Chronic Lymphocytic Leukemia

N.K. Yankovsky
N.I. Vavilov Institute of General Genetics, Moscow, Russia

Informatics

BCM Server Core

Daniel Davison
Baylor College of Medicine, Houston, Texas

A Freely Sharable Database-Management System Designed for Use in Component-Based, Modular Genome Informatics Systems

Nathan Goodman
The Jackson Laboratory, Bar Harbor, Maine

A Software Environment for Large-Scale Sequencing

Mark Graves
Baylor College of Medicine, Houston, Texas

Generalized Hidden Markov Models for Genomic Sequence Analysis

David Haussler
University of California, Santa Cruz

Identification, Organization, and Analysis of Mammalian Repetitive DNA Information

Jerzy Jurka
Genetic Information Research Institute, Palo Alto, California

*TRRD, GERD and COMPEL: Databases on Gene-Expression Regulation as a Tool for Analysis of Functional Genomic Sequences

N.A. Kolchanov
Institute of Cytology and Genetics, Novosibirsk, Russia

Data-Management Tools for Genomic Databases

Victor M. Markowitz and I-Min A. Chen
Lawrence Berkeley National Laboratory, Berkeley, California

The Genome Topographer: System Design

T. Marr
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

A Flexible Sequence Reconstructor for Large-Scale DNA Sequencing: A Customizable Software System for Fragment Assembly

Gene Myers
University of Arizona, Tucson

The Role of Integrated Software and Databases in Genome Sequence Interpretation and Metabolic Reconstruction

Ross Overbeek
Argonne National Laboratory, Argonne, Illinois

Database Transformations for Biological Applications

G. Christian Overton, Susan B. Davidson, and **Peter Buneman**
University of Pennsylvania, Philadelphia

Las Vegas Algorithm for Gene Recognition: Suboptimal and Error-Tolerant Spliced Alignment

Pavel A. Pevzner
University of Southern California, Los Angeles, California

Foundations for a Syntactic Pattern-Recognition System for Genomic DNA Sequences: Languages, Automata, Interfaces, and Macromolecules

David B. Searls
SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania

Analysis and Annotation of Nucleic Acid Sequence

David J. States
Washington University, St. Louis, Missouri

Gene Recognition, Modeling, and Homology Search in GRAIL and genQuest

Edward C. Uberbacher
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Informatics Support for Mapping in Mouse-Human Homology Regions

Edward Uberbacher
Oak Ridge National Laboratory, Oak Ridge, Tennessee

SubmitData: Data Submission to Public Genomic Databases

Manfred D. Zorn
Lawrence Berkeley National Laboratory, University of California, Berkeley

ELSI

The Human Genome: Science and the Social Consequences; Interactive Exhibits and Programs on Genetics and the Human Genome

Charles C. Carlson
The Exploratorium, San Francisco, California

Documentary Series for Public Broadcasting

Graham Chedd and **Noel Schwerin**
Chedd-Angier Production Company, Watertown, Massachusetts

Human Genome Teacher Networking Project

Debra L. Collins and **R. Neil Schimke**
University of Kansas Medical Center, Kansas City, Kansas

Human Genome Education Program

Lane Conn
Stanford Human Genome Center, Palo Alto, California

Your World/Our World–Biotechnology & You: Special Issue on the Human Genome Project

Jeff Davidson and **Laurence Weinberger**
Pennsylvania Biotechnology Association, State College, Pennsylvania

The Human Genome Project and Mental Retardation: An Educational Program

Sharon Davis
The Arc of the United States, Arlington, Texas

Pathways to Genetic Screening: Molecular Genetics Meets the High-Risk Family

Troy Duster
University of California, Berkeley

Intellectual Property Issues in Genomics

Rebecca S. Eisenberg
University of Michigan Law School, Ann Arbor, Michigan

AAAS Congressional Fellowship Program

Stephen Goodman
The American Society of Human Genetics, Bethesda, Maryland

A Hispanic Educational Program for Scientific, Ethical, Legal, and Social Aspects of the Human Genome Project

Margaret C. Jefferson and **Mary Ann Sesma**[1]
California State University and [1]Los Angeles Unified School District, Los Angeles, California

Implications of the Geneticization of Health Care for Primary Care Practitioners

Mary B. Mahowald
University of Chicago, Chicago, Illinois

Nontraditional Inheritance: Genetics and the Nature of Science; Instructional Materials for High School Biology

Joseph D. McInerney and **B. Ellen Friedman**
Biological Sciences Curriculum Study, Colorado Springs, Colorado

The Human Genome Project: Biology, Computers, and Privacy: Development of Educational Materials for High School Biology

Joseph D. McInerney and **Lynda B. Micikas**
Biological Sciences Curriculum Study, Colorado Springs, Colorado

Involvement of High School Students in Sequencing the Human Genome

Maureen M. Munn, **Maynard V. Olson**, and **Leroy Hood**
University of Washington, Seattle

The Gene Letter: A Newsletter on Ethical, Legal, and Social Issues in Genetics for Interested Professionals and Consumers

Philip J. Reilly, **Dorothy C. Wertz**, and **Robin J.R. Blatt**
The Shriver Center for Mental Retardation, Waltham, Massachusetts

The DNA Files: A Nationally Syndicated Series of Radio Programs on the Social Implications of Human Genome Research and Its Applications

Bari Scott
Genome Radio Project, KPFA-FM, Berkeley, California

Communicating Science in Plain Language: The Science+ Literacy for Health: Human Genome Project

Maria Sosa, Judy Kass, and Tracy Gath
American Association for the Advancement of Science, Washington, D.C.

The Community College Initiative

Sylvia J. Spengler and Laurel Egenberger
Lawrence Berkeley National Laboratory, Berkeley, California

Genome Educators

Sylvia Spengler and **Janice Mann**
Lawrence Berkeley National Laboratory, Berkeley, California

Getting the Word Out on the Human Genome Project: A Course for Physicians

Sara L. Tobin and **Ann Boughton**'
Stanford University, Palo Alto, California
'Thumbnail Graphics, Oklahoma City, Oklahoma

The Genetics Adjudication Resource Project

Franklin M. Zweig
Einstein Institute for Science, Health, and the Courts, Bethesda, Maryland

Infrastructure

Alexander Hollaender Distinguished Postdoctoral Fellowships

Linda Holmes and **Eugene Spejewski**
Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

Human Genome Management Information System

Betty K. Mansfield and **John S. Wassom**
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Human Genome Program Coordination

Sylvia J. Spengler
Lawrence Berkeley National Laboratory, Berkeley, California

Support of Human Genome Program Proposal Reviews

Walter Williams
Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

Former Soviet Union Office of Health and Environmental Research Program

James Wright
Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

SBIR

1996 Phase I

An Engineered RNA/DNA Polymerase to Increase Speed and Economy of DNA Sequencing

Mark W. Knuth
Promega Corporation, Madison, Wisconsin

Directed Multiple DNA Sequencing and Expression Analysis by Hybridization

Gualberto Ruano
BIOS Laboratories, Inc., New Haven, Connecticut

1996 Phase II

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogeneous Public Genome Databases

Joseph Leone
CyberConnect Corporation, Storrs, Connecticut

Low-Cost Automated Preparation of Plasmid, Cosmid, and Yeast DNA

William P. MacConnell
MacConnell Research Corporation, San Diego, California

GRAIL-GenQuest: A Comprehensive Computational Framework for DNA Sequence Analysis

Ruth Ann Manning
ApoCom, Inc., Oak Ridge, Tennessee

Appendix F: DOE BER Program

Text and photos in this appendix first appeared in a brochure prepared by the Human Genome Management Information System for the DOE Office of Biological and Environmental Research to announce a symposium celebrating 50 years of achievements in the Biological and Environmental Research Program. "Serving Science and Society into the New Millennium" was held on May 21–22, 1997, at the National Academy of Sciences in Washington, D.C. The color brochure and other recent publications related to BER research, including the historically comprehensive A Vital Legacy, *may be obtained from HGMIS at the address on the inside front cover.*

DOE Biological and Environmental Research Program

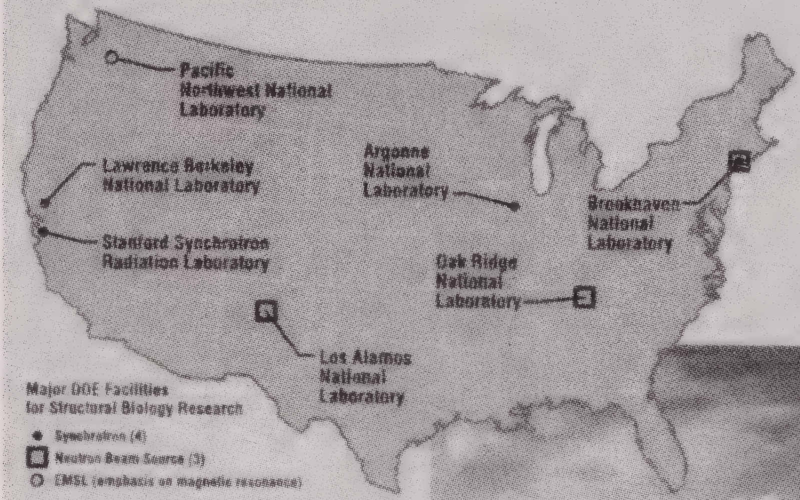
An Extraordinary Legacy

To exploit the boundless promise of energy technologies and shed light on their consequences to public health and the environment, the Biological and Environmental Research program of the U.S. Department of Energy's (DOE) Office of Health and Environmental Research (OHER) has engaged in a variety of multidisciplinary research activities:

- Establishing the world's first Human Genome Program.
- Developing advanced medical diagnostic tools and treatments for human disease.
- Assessing the health effects of radiation.

BIOTA TO GLOBAL CHANGE • MOLECULES TO HUMANS

Biological and Environmental Research Program
Aristides Patrinos, Ph.D.
Associate Director for Energy Research
for the
Office of Biological and Environmental Research
U.S. Department of Energy
301/903-3251, Fax: 301/903-5051
http://www.er.doe.gov/production/ober/ober_top.html


Pacific Northwest National Laboratory
Lawrence Berkeley National Laboratory
Argonne National Laboratory
Brookhaven National Laboratory
Stanford Synchrotron Radiation Laboratory
Oak Ridge National Laboratory
Los Alamos National Laboratory
Major DOE Facilities for Structural Biology Research
● Synchrotron (4)
☐ Neutron Beam Source (3)
○ EMSL (emphasis on magnetic resonance)

National User Facilities

Dedicated biomedical resources, such as those maintained by BER at several DOE laboratories, are available at little or no charge. These resources enable scientists to gain an understanding of relationships between biological structures and their functions, study disease processes, develop new pharmaceuticals, and conduct basic research in molecular biology and environmental processes.

William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) is a national collaborative user facility for providing innovative approaches to meet the needs of DOE's environmental missions.

An Enduring Mandate

DOE is carrying forward Congressional mandates that began with its predecessors, the Atomic Energy Commission and the Energy Research and Development Agency:

Contribute to a Healthy Citizenry

- Develop innovative technologies for tomorrow's biomedical sciences.
- Provide the basis for individual risk assessments by determining the human genome's fine structure by the year 2005.
- Conduct research into advanced medical technologies and radiopharmaceuticals.
- Build and support national user facilities for determining biological structure, and ultimately function, at the molecular and cellular level.

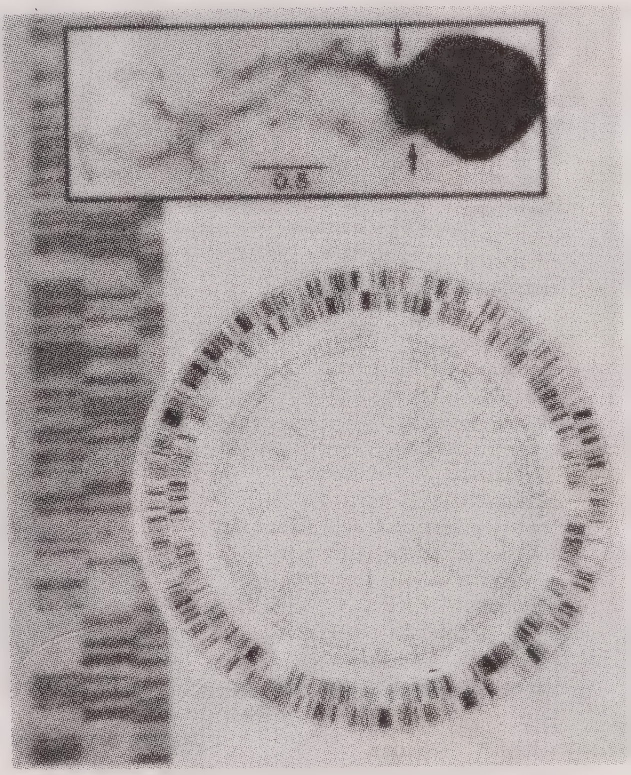
Understand Global Climate Change

Predict the effects of energy production and its use on the regional and global environment by acquiring data and developing the necessary understanding of environmental processes.

Contribute to Environmental Cleanup

Conduct fundamental research to establish a better scientific basis for remediating contaminated sites.

DOE user facilities are revealing the molecular details of life. Knowing the 3-D structure of the ras protein (above), an important molecular switch governing human cell growth, will enable interventions to shut off this switch in cancer cells.



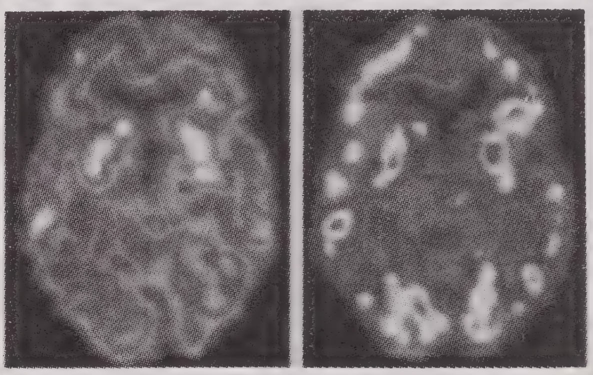
Determining the fine structure—DNA sequence—of the microorganism *Methanococcus jannaschii* (pictured at right, top) and other minimal life forms in DOE's Microbial Genome Program will benefit medicine, agriculture, industrial and energy production, and environmental bioremediation. The circular representation of the single *M. jannaschii* chromosome, which was fully sequenced in 1996, illustrates the location of genes and other important features. (Vertical bar represents a portion of a sequencing experiment.)

Fifty Years of Achievements. . . Leading to Innovative Solutions

Tools for Medicine and Research

Radioisotopes developed for medicine and medical imaging are being merged with current knowledge in biology and genetics to discover new ways of diagnosing and treating cancer and other disorders, detecting genes in action, and understanding normal development and function of human organ systems.

- Radioactive molecules used in medical imaging for positron emission tomography (PET) and magnetic resonance imaging (MRI) allow noninvasive diagnosis, monitoring, and exploration of human disorders and their treatments.
- Isotopes and other tracers of brain activity are being used to explore drug addiction, the effects of smoking, Alzheimer's disease, Parkinson's disease, and schizophrenia.
- Technetium-99m is used to diagnose diseases of the kidney, liver, heart, brain, and other organs in about 13 million patients per year.
- Striking successes have been achieved using charged atomic particles to treat thyroid diseases, pituitary tumors, and eye cancer, among other disorders.



One-quarter of all patients in U.S. hospitals undergo tests using descendants of cameras developed by BER to follow radioactive tracers in the body. PET scanning has been key to a generation of brain metabolism studies as well as diagnostic tests for heart disease and cancer. PET studies above reveal brain metabolism differences in recovering alcoholics (left, 10 days, and right, 30 days, after withdrawal from alcohol).

Genome Projects

A legacy of DOE research on genetic effects paved the way for the world's first Human Genome Program. Now new genomic technologies are being applied to environmental cleanup through the DOE Natural and Accelerated Bioremediation Research and Microbial Genome programs, healthcare and risk assessment, and such other national priorities as industrial processes and agriculture.

The laser-based flow cytometer developed at DOE national laboratories enables researchers to separate human chromosomes for analysis.

Discover the breadth of current activities and recent accomplishments via the BER Web Site:

http://www.er.doe.gov/production/ober/ober_top.html

Radiation Risks and Protection Guidelines

BER studies have become the foundation for laws and standards that protect the population, including workers exposed to radiological sources:

- Guidelines for the safe use of diagnostic X rays and radiopharmaceuticals.
- Safety standards for the presence of radionuclides in food and drinking water.
- Radiation-detection systems and dosimetry techniques.

Finding a Link Between DNA Damage and Cancers

Studies of DNA damage have uncovered similar mechanisms at work in damage caused by radiation exposure, X rays, ultraviolet light, and cancer-causing chemicals. A screening test for such chemicals is now one of the first hurdles a new compound must clear on its way to regulatory and public acceptance.

Tracking the Regional and Global Movement of Pollutants

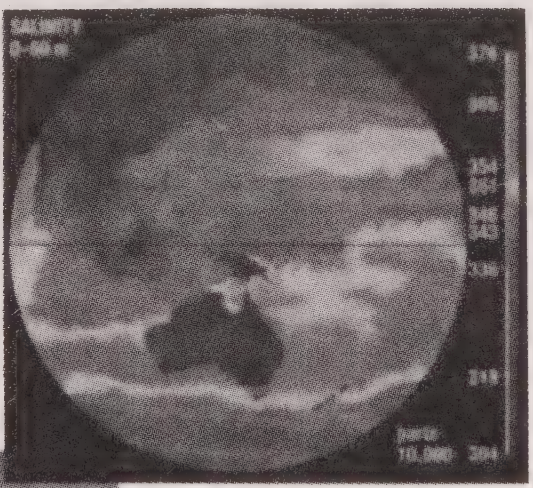
BER research helped to establish the earliest and most authoritative monitoring network in the world to detect airborne radioisotopes. The use of atmospheric tracers has led to the improved ability to predict the dispersion of pollutants.

Understanding Global Change

Important achievements in environmental research have led to enhanced capabilities in studying global change, including more accurate predictions of global and regional climate changes induced by increasing atmospheric concentrations of greenhouse gases.

Human chromosomes "painted" by fluorescent dyes to detect abnormal exchange of genetic material frequently present in cancer. Chromosome paints also serve as valuable resources for other clinical and research applications.

> ***". . . (it's) not so much where we stand as in what direction we are moving. [Oliver Wendell Homes, Sr.]"***



High-performance computing is promoting faster and more realistic solutions to long-term climate change.

The Unmanned Aerospace Vehicle (above) conducts measurements to quantify the fate of solar radiation falling on the earth.

Creating a New Science of Ecology

BER achievements in using radioactive tracers to follow the movements of animals, routes of chemicals through food chains, decomposition of forest detritus, together with the program's introduction of computer simulations, created the new field of radioecology.

Glossary

This glossary was adapted from definitions in the DOE *Primer on Molecular Genetics* (1992).

http://www.ornl.gov/hgmis/publicat/primer/[illegible]

A

Adenine (A): A nitrogenous base, one member of the base pair A-T (adenine-thymine).

Allele: Alternative form of a genetic locus; a single allele for each locus is inherited separately from each parent (e.g., at a locus for eye color the allele might result in blue or brown eyes).

Amino acid: Any of a class of 20 molecules that are combined to form proteins in living things. The sequence of amino acids in a protein and hence protein function are determined by the genetic code.

Amplification: An increase in the number of copies of a specific DNA fragment; can be in vivo or in vitro. See cloning, polymerase chain reaction.

Arrayed library: Individual primary recombinant clones (hosted in phage, cosmid, YAC, or other vector) that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified by the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications, including screening for a specific gene or genomic region of interest as well as for physical mapping. Information gathered on individual clones from various genetic linkage and physical map analyses is entered into a relational database and used to construct physical and genetic linkage maps simultaneously; clone identifiers serve to interrelate the multilevel maps. Compare library, genomic library.

Autoradiography: A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel electrophoresis.

Autosome: A chromosome not involved in sex determination. The diploid human genome consists of 46 chromosomes, 22 pairs of autosomes, and 1 pair of sex chromosomes (the X and Y chromosomes).

B

BAC: See bacterial artificial chromosome.

Bacterial artificial chromosome (BAC): A vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on naturally occurring F-factor plasmid found in the bacterium *E. coli*. Compare cloning vector.

Bacteriophage: See phage.

Base pair (bp): Two nitrogenous bases (adenine and thymine or guanine and cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Base sequence: The order of nucleotide bases in a DNA molecule.

Base sequence analysis: A method, sometimes automated, for determining the base sequence.

Biotechnology: A set of biological techniques developed through basic research and now applied to research and product development. In particular, the use by industry of recombinant DNA, cell fusion, and new bioprocessing techniques.

bp: See base pair.

C

cDNA: See complementary DNA.

Centimorgan (cM): A unit of measure of recombination frequency. One centimorgan is equal to a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million base pairs.

Centromere: A specialized chromosome region to which spindle fibers attach during cell division.

Chromosome: The self-replicating genetic structure of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins.

Clone bank: See genomic library.

Clone: A group of cells derived from a single ancestor.

Cloning: The process of asexually producing a group of cells (clones), all genetically identical, from a single ancestor. In recombinant DNA technology, the use of DNA manipulation procedures to produce multiple copies of a single gene or segment of DNA is referred to as cloning DNA.

Cloning vector: DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vectors capacity for self-replication; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artificial chromosomes; vectors are often recombinant molecules containing DNA sequences from several sources.

cM: See centimorgan.

Code: See genetic code.

Codon: See genetic code.

Complementary DNA (cDNA): DNA that is synthesized from a messenger RNA template; the single-stranded form is often used as a probe in physical mapping.

Complementary sequence: Nucleic acid base sequence that can form a double-stranded structure by matching base pairs with another sequence; the complementary sequence to G-T-A-C is C-A-T-G.

Conserved sequence: A base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged throughout evolution.

Contig: Group of clones representing overlapping regions of a genome.

Contig map: A map depicting the relative order of a linked library of small overlapping clones representing a complete chromosomal segment.

Cosmid: Artificially constructed cloning vector containing the cos gene of phage lambda. Cosmids can be packaged in lambda phage particles for infection into *E. coli*; this permits cloning of larger DNA fragments (up to 45 kb) than can be introduced into bacterial hosts in plasmid vectors.

Crossing over: The breaking during meiosis of one maternal and one paternal chromosome, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of alleles between chromosomes. Compare recombination.

Cytosine (C): A nitrogenous base, one member of the base pair G-C (guanine and cytosine).

D

Deoxyribonucleotide: See nucleotide.

Diploid: A full set of genetic material, consisting of paired chromosomes one chromosome from each parental set. Most animal cells except the gametes have a diploid set of chromosomes. The diploid human genome has 46 chromosomes. Compare haploid.

DNA (deoxyribonucleic acid): The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus the base sequence of each single strand can be deduced from that of its partner.

DNA probe: See probe.

DNA replication: The use of existing DNA as a template for the synthesis of new DNA strands. In humans and other eukaryotes, replication occurs in the cell nucleus.

DNA sequence: The relative order of base pairs, whether in a fragment of DNA, a gene, a chromosome, or an entire genome. See base sequence analysis.

Domain: A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function.

Double helix: The shape that two linear strands of DNA assume when bonded together.

E

E. coli: Common bacterium that has been studied intensively by geneticists because of its small genome size, normal lack of pathogenicity, and ease of growth in the laboratory.

Electrophoresis: A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Separation is based on these differences. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids.

Endonuclease: An enzyme that cleaves its nucleic acid substrate at internal sites in the nucleotide sequence.

Enzyme: A protein that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction.

EST: Expressed sequence tag. See sequence tagged site.

Eukaryote: Cell or organism with membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and blue-green algae. Compare prokaryote. See chromosome.

Evolutionarily conserved: See conserved sequence.

Exogenous DNA: DNA originating outside an organism.

Exon: The protein-coding DNA sequence of a gene. Compare intron.

Exonuclease: An enzyme that cleaves nucleotides sequentially from free ends of a linear nucleic acid substrate.

Expressed gene: See gene expression.

F

FISH (fluorescence in situ hybridization): A physical mapping approach that uses fluorescein tags to detect hybridization of probes with metaphase chromosomes and with the less-condensed somatic interphase chromatin.

Flow cytometry: Analysis of biological material by detection of the light-absorbing or fluorescing properties of cells or subcellular fractions (i.e., chromosomes) passing in a narrow stream through a laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, sort successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet.

Flow karyotyping: Use of flow cytometry to analyze and separate chromosomes on the basis of their DNA content.

G

Gamete: Mature male or female reproductive cell (sperm or ovum) with a haploid set of chromosomes (23 for humans).

Gene: The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule). See gene expression.

Gene expression: The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs).

Gene family: Group of closely related genes that make similar products.

Gene library: See genomic library.

Gene mapping: Determination of the relative positions of genes on a DNA molecule (chromosome or plasmid) and of the distance, in linkage units or physical units, between them.

Gene product: The biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease-causing alleles.

Genetic code: The sequence of nucleotides, coded in triplets (codons) along the mRNA, that determines the sequence of amino acids in protein synthesis. The DNA sequence of a gene can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence.

Genetic engineering technology: See recombinant DNA technology.

Genetic map: See linkage map.

Genetic material: See genome.

Genetics: The study of the patterns of inheritance of specific traits.

Genome: All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.

Genome project: Research and technology development effort aimed at mapping and sequencing some or all of the genome of human beings and other organisms.

Genomic library: A collection of clones made from a set of randomly generated overlapping DNA fragments representing the entire genome of an organism. Compare library, arrayed library.

Guanine (G): A nitrogenous base, one member of the base pair G-C (guanine and cytosine).

H

Haploid: A single set of chromosomes (half the full set of genetic material), present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. Compare diploid.

Heterozygosity: The presence of different alleles at one or more loci on homologous chromosomes.

Homeobox: A short stretch of nucleotides whose base sequence is virtually identical in all the genes that contain it. It has been found in many organisms from fruit flies to human beings. In the fruit fly, a homeobox appears to determine when particular groups of genes are expressed during development.

Homology: Similarity in DNA or protein sequences between individuals of the same species or among different species.

Homologous chromosome: Chromosome containing the same linear gene sequences as another, each derived from one parent.

Human gene therapy: Insertion of normal DNA directly into cells to correct a genetic defect.

Human Genome Initiative: Collective name for several projects begun in 1986 by DOE to (1) create an ordered set of DNA segments from known chromosomal locations, (2) develop new computational methods for analyzing genetic map and DNA sequence data, and (3) develop new techniques and instruments for detecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The national effort, led by DOE and NIH, is known as the Human Genome Project.

Hybridization: The process of joining two complementary strands of DNA or one each of DNA and RNA to form a double-stranded molecule.

I

Informatics: The study of the application of computer and statistical techniques to the management of information. In genome projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence information, and to predict protein sequence and structure from DNA sequence data.

In situ hybridization: Use of a DNA or RNA probe to detect the presence of the complementary DNA sequence in cloned bacterial or cultured eukaryotic cells.

Interphase: The period in the cell cycle when DNA is replicated in the nucleus; followed by mitosis.

Intron: The DNA base sequence interrupting the protein-coding sequence of a gene; this sequence is transcribed into RNA but is cut out of the message before it is translated into protein. Compare exon.

In vitro: Outside a living organism.

K

Karyotype: A photomicrograph of an individual's chromosomes arranged in a standard format showing the number, size, and shape of each chromosome type; used in low-resolution physical mapping to correlate gross chromosomal abnormalities with the characteristics of specific diseases.

kb: See kilobase.

Kilobase (kb): Unit of length for DNA fragments equal to 1000 nucleotides.

L

Library: An unordered collection of clones (i.e., cloned DNA from a particular organism), whose relationship to each other can be established by physical mapping. Compare genomic library, arrayed library.

Linkage: The proximity of two or more markers (e.g., genes, RFLP markers) on a chromosome; the closer together the markers are, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together.

Linkage map: A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans (cM).

Localize: Determination of the original position (locus) of a gene or other marker on a chromosome.

Locus (pl. loci): The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position. The use of locus is sometimes restricted to mean regions of DNA that are expressed. See gene expression.

M

Macrorestriction map: Map depicting the order of and distance between sites at which restriction enzymes cleave chromosomes.

Mapping: See gene mapping, linkage map, physical map.

Marker: An identifiable physical location on a chromosome (e.g., restriction enzyme cutting site, gene) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined. See RFLP, restriction fragment length polymorphism.

Mb: See megabase.

Megabase (Mb): Unit of length for DNA fragments equal to 1 million nucleotides and roughly equal to 1 cM.

Meiosis: The process of two consecutive cell divisions in the diploid progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a haploid set of chromosomes.

Messenger RNA (mRNA): RNA that serves as a template for protein synthesis. See genetic code.

Metaphase: A stage in mitosis or meiosis during which the chromosomes are aligned along the equatorial plane of the cell.

Mitosis: The process of nuclear division in cells that produces daughter cells that are genetically identical to each other and to the parent cell.

mRNA: See messenger RNA.

Multifactorial or multigenic disorder: See polygenic disorder.

Multiplexing: A sequencing approach that uses several pooled samples simultaneously, greatly increasing sequencing speed.

Mutation: Any heritable change in DNA sequence. Compare polymorphism.

N

Nitrogenous base: A nitrogen-containing molecule having the chemical properties of a base.

Nucleic acid: A large molecule composed of nucleotide subunits.

Nucleotide: A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule. See DNA, base pair, RNA.

Nucleus: The cellular organelle in eukaryotes that contains the genetic material.

O

Oncogene: A gene, one or more forms of which is associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth.

Overlapping clones: See genomic library.

P

P1-derived artificial chromosome (PAC): A vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on bacteriophage (a virus) P1 genome. Compare cloning vector.

PAC: See P1-derived artificial chromosome.

PCR: See polymerase chain reaction.

Phage: A virus for which the natural host is a bacterial cell.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest-resolution physical map is the banding patterns on the 24 different chromosomes; the highest-resolution map would be the complete nucleotide sequence of the chromosomes.

Plasmid: Autonomously replicating, extrachromosomal circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors.

Polygenic disorder: Genetic disorder resulting from the combined action of alleles of more than one gene (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles; thus the hereditary patterns are usually more complex than those of single-gene disorders. Compare single-gene disorders.

Polymerase chain reaction (PCR): A method for amplifying a DNA base sequence using a heat-stable polymerase and two 20-base primers, one complementary to the (+)-strand at one end of the sequence to be amplified and the other complementary to the (-)-strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample.

Polymerase, DNA or RNA: Enzymes that catalyze the synthesis of nucleic acids on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides.

Polymorphism: Difference in DNA sequence among individuals. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic linkage analysis. Compare mutation.

Primer: Short preexisting polynucleotide chain to which new deoxyribonucleotides can be added by DNA polymerase.

Probe: Single-stranded DNA or RNA molecules of specific base sequence, labeled either radioactively or immunologically, that are used to detect the complementary base sequence by hybridization.

Prokaryote: Cell or organism lacking a membrane-bound, structurally discrete nucleus and other subcellular compartments. Bacteria are prokaryotes. Compare eukaryote. See chromosome.

Promoter: A site on DNA to which RNA polymerase will bind and initiate transcription.

Protein: A large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, function, and regulation of the bodys cells, tissues, and organs, and each protein has unique functions. Examples are hormones, enzymes, and antibodies.

Purine: A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine.

Pyrimidine: A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine; in RNA, cytosine and uracil.

R

Rare-cutter enzyme: See restriction enzyme cutting site.

Recombinant clone: Clone containing recombinant DNA molecules. See recombinant DNA technology.

Recombinant DNA molecules: A combination of DNA molecules of different origin that are joined using recombinant DNA technologies.

Recombinant DNA technology: Procedure used to join together DNA segments in a cell-free system (an environment outside a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular chromosome.

Recombination: The process by which progeny derive a combination of genes different from that of either parent. In higher organisms, this can occur by crossing over.

Regulatory region or sequence: A DNA base sequence that controls gene expression.

Resolution: Degree of molecular detail on a physical map of DNA, ranging from low to high.

Restriction enzyme, endonuclease: A protein that recognizes specific, short nucleotide sequences and cuts DNA at those sites. Bacteria contain over 400 such enzymes that recognize and cut over 100 different DNA sequences. See restriction enzyme cutting site.

Restriction enzyme cutting site: A specific nucleotide sequence of DNA at which a particular restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs), others much less frequently (rare-cutter; e.g., every 10,000 base pairs).

Restriction fragment length polymorphism (RFLP): Variation between individuals in DNA fragment sizes cut by specific restriction enzymes; polymorphic sequences that result in RFLPs are used as markers on both physical maps and genetic linkage maps. RFLPs are usually caused by mutation at a cutting site. See marker.

RFLP: See restriction fragment length polymorphism.

Ribonucleic acid (RNA): A chemical found in the nucleus and cytoplasm of cells; it plays an important role in protein synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including messenger RNA, transfer RNA, ribosomal RNA, and other small RNAs, each serving a different purpose.

Ribonucleotide: See nucleotide.

Ribosomal RNA (rRNA): A class of RNA found in the ribosomes of cells.

Ribosomes: Small cellular components composed of specialized ribosomal RNA and protein; site of protein synthesis. See ribonucleic acid (RNA).

RNA: See ribonucleic acid.

S

Sequence: See base sequence.

Sequence tagged site (STS): Short (200 to 500 base pairs) DNA sequence that has a single occurrence in the human genome and whose location and base sequence are known. Detectable by polymerase chain reaction, STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and serve as landmarks on the developing physical map of the human genome. Expressed sequence tags (ESTs) are STSs derived from cDNAs.

Sequencing: Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

Sex chromosome: The X or Y chromosome in human beings that determines the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a karyotype. Compare autosome.

Shotgun method: Cloning of DNA fragments randomly generated from a genome. See library, genomic library.

Single-gene disorder: Hereditary disorder caused by a mutant allele of a single gene (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). Compare polygenic disorders.

Somatic cell: Any cell in the body except gametes and their precursors.

Southern blotting: Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific base sequences by radiolabeled complementary probes.

STS: See sequence tagged site.

T

Tandem repeat sequences: Multiple copies of the same base sequence on a chromosome; used as a marker in physical mapping.

Technology transfer: The process of converting scientific findings from research laboratories into useful products by the commercial sector.

Telomere: The end of a chromosome. This specialized structure is involved in the replication and stability of linear DNA molecules. See DNA replication.

Thymine (T): A nitrogenous base, one member of the base pair A-T (adenine-thymine).

Transcription: The synthesis of an RNA copy from a sequence of DNA (a gene); the first step in gene expression. Compare translation.

Transfer RNA (tRNA): A class of RNA having structures with triplet nucleotide sequences that are complementary to the triplet nucleotide coding sequences of mRNA. The role of tRNAs in protein synthesis is to bond with amino acids and transfer them to the ribosomes, where proteins are assembled according to the genetic code carried by mRNA.

Transformation: A process by which the genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its genome.

Translation: The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. Compare transcription.

tRNA: See transfer RNA.

U

Uracil: A nitrogenous base normally found in RNA but not DNA; uracil is capable of forming a base pair with adenine.

V

Vector: See cloning vector.

Virus: A noncellular biological entity that can reproduce only within a host cell. Viruses consist of nucleic acid covered by protein; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus.

VLSI: Very large scale integration allowing more than 100,000 transistors on a chip.

Y

YAC: See yeast artificial chromosome.

Yeast artificial chromosome (**YAC**)**:** A vector used to clone DNA fragments (up to 400 kb); it is constructed from the telomeric, centromeric, and replication origin sequences needed for replication in yeast cells. Compare cloning vector

DOE/ER-0713 (Part 2)

HUMAN GENOME PROGRAM REPORT

Part 2, 1996 Research Abstracts

Date Published: November 1997

Prepared for the
U.S. Department of Energy
Office of Energy Research
Office of Biological and Environmental Research
Germantown, MD 20874-1290

Prepared by the
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830-6480
managed by
Lockheed Martin Energy Research Corporation
for the
U.S. Department of Energy
Under Contract DE-AC05-96OR22464

Preface

More than a decade ago, the Office of Health and Environmental Research (OHER) of the U.S. Department of Energy (DOE) struck a bold course in launching its Human Genome Initiative, convinced that its mission would be well served by a comprehensive picture of the human genome. Organizers recognized that the information the project would generate—both technological and genetic—would contribute not only to a new understanding of human biology and the effects of energy technologies but also to a host of practical applications in the biotechnology industry and in the arenas of agriculture and environmental protection.

Today, the project's value appears beyond doubt as worldwide participation contributes toward the goals of determining the human genome's complete sequence by 2005 and elucidating the genome structure of several model organisms as well. This report summarizes the content and progress of the DOE Human Genome Program (HGP). Descriptive research summaries, along with information on program history, goals, management, and current research highlights, provide a comprehensive view of the DOE program.

Last year marked an early transition to the third and final phase of the U.S. Human Genome Project as pilot programs to refine large-scale sequencing strategies and resources were funded by DOE and the National Institutes of Health, the two sponsoring U.S. agencies. The human genome centers at Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory had been serving as the core of DOE multidisciplinary HGP research, which requires extensive contributions from biologists, engineers, chemists, computer scientists, and mathematicians. These team efforts were complemented by those at other DOE-supported laboratories and about 60 universities, research organizations, companies, and foreign institutions. Now, to focus DOE's considerable resources on meeting the challenges of large-scale sequencing, the sequencing efforts of the three genome centers have been integrated into the Joint Genome Institute. The institute will continue to bring together research from other DOE-supported laboratories. Work in other critical areas continues to develop the resources and technologies needed for production sequencing; computational approaches to data management and interpretation (called informatics); and an exploration of the important ethical, legal, and social issues arising from use of the generated data, particularly regarding the privacy and confidentiality of genetic information.

Insights, technologies, and infrastructure emerging from the Human Genome Project are catalyzing a biological revolution. Health-related biotechnology is already a success story—and is still far from reaching its potential. Other applications are likely to beget similar successes in coming decades; among these are several of great importance to DOE. We can look to improvements in waste control and an exciting era of environmental bioremediation, we will see new approaches to improving energy efficiency, and we can hope for dramatic strides toward meeting the fuel demands of the future.

In 1997 OHER, renamed the Office of Biological and Environmental Research (OBER), is celebrating 50 years of conducting research to exploit the boundless promise of energy technologies while exploring their consequences to the public's health and the environment. The DOE Human Genome Program and a related spin-off project, the Microbial Genome Program, are major components of the Biological and Environmental Research Program of OBER.

DOE OBER is proud of its contributions to the Human Genome Project and welcomes general or scientific inquiries concerning its genome programs. Announcements soliciting research applications appear in *Federal Register*, *Science*, *Human Genome News*, and other publications. The deadline for formal applications is generally midsummer for awards to be made the next year, and submission of preproposals in areas of potential interest is strongly encouraged. Further information may be obtained by contacting the program office or visiting the DOE home page (301/903-6488, Fax: -8521, *genome@oer.doe.gov*, URL: *http://www.er.doe.gov/production/ober/hug_top.html*).

Aristides Patrinos, Associate Director
Office of Biological and Environmental Research
U.S. Department of Energy
November 3, 1997

Foreword

The research abstracts in this section were funded in FY 1996 by the DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997.

These unedited abstracts were contributed by DOE Human Genome Program grantees and contractors. Names of principal investigators are in bold print. Submitted in 1996, contact information is for the first person named unless another investigator is designated as contact person. Principal investigators of research projects described by abstracts in this section are listed under their respective subject categories, and an index of all investigators named in the abstracts is given at the end of this report.

Part 1 of this report contains narratives that represent DOE Human Genome Program research in large, multidisciplinary projects. As a convenience to the reader, these narratives are reprinted (without graphics) as an appendix to this volume, Part 2. The projects represent work at the Joint Genome Institute (p. 72), Lawrence Livermore National Laboratory Human Genome Center (p. 73), Los Alamos National Laboratory Center for Human Genome Studies (p. 77), Lawrence Berkeley National Laboratory Human Genome Center (p. 81), University of Washington Genome Center (p. 85), Genome Database (p. 87), and National Center for Genome Resources (p. 91). Only the contact persons for these organizations are listed in the Index to Principal and Coinvestigators. More information on research carried out in these projects can be found on their listed Web sites.

Contents

1996 Research Abstracts 1

Sequencing 1

Mapping 19

Informatics 33

Ethical, Legal, and Social Issues 45

Infrastructure 59

Small Business Innovative Research 63

Projects Completed FY 1994–95 67

Appendix: Narratives from Large, Multidisciplinary Research Projects 71

(Text reprinted from *Human Genome Program Report: Part 1, Overview and Progress*)

Index to Principal and Coinvestigators 93

Acronym List ..... Inside back cover

1996 Research Abstracts

Project Categories and Principal Investigators

Sequencing 1
- W.H. Benner and J.M. Jaklevic 1
- Chung-Hsuan Chen 1
- George Church 2
- Pieter de Jong 2
- Norman Dovichi 3
- John J. Dunn and F. William Studier 3
- John J. Dunn and F. William Studier 4
- Glen A. Evans 4
- Glen A. Evans 5
- *V.L. Florentiev 5
- Raymond F. Gesteland 6
- Joe Gray and Daniel Pinkel 7
- Trevor Hawkins 8
- Leroy Hood, Mark D. Adams, and Melvin Simon 8
- Barry L. Karger 9
- Richard A. Mathies and Alexander N. Glazer 9
- Andrei Mirzabekov 10
- Robert K. Moyzis 12
- Barbara Ramsay Shaw 13
- Lloyd M. Smith and Richard A. Guilfoyle 13
- Lloyd M. Smith 14
- Richard D. Smith 14
- Richard D. Smith 15
- Stanley Tabor 16
- Levy Ulanovsky 16
- Peter Williams 17
- Edward S. Yeung 17

Mapping 19
- David Allison and Bruce Warmack 19
- *Alexandre S. Boitsov 19
- Charles R. Cantor 19
- Pieter J. de Jong 20
- Jacques R. Fresco 21
- Fa-Ten Kao 21
- *V.L. Karpov 22

---

*Russian scientists designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).

Julie R. Korenberg 22
D. L. Nelson 23
***Olga Podgornaya** 24
***O.L. Polanovsky** 25
***A.I. Poletaev** 26
Melvin I. Simon 26
Melvin I. Simon 27
Marcelo Bento Soares 27
Lisa Stubbs 28
Lisa Stubbs 29
Jean-Michel H. Vos 30
***N.K. Yankovsky** 30

Informatics 33
Daniel Davison 33
Nathan Goodman 33
Mark Graves 34
David Haussler 34
Jerzy Jurka 34
***N.A. Kolchanov** 35
Victor M. Markowitz and I-Min A. Chen 36
T. Marr 37
Gene Myers 38
Ross Overbeek 38
G. Christian Overton, Susan B. Davidson, and Peter Buneman 39
Pavel A. Pevzner 40
David B. Searls 41
David J. States 41
Edward C. Uberbacher 42
Edward Uberbacher 44
Manfred D. Zorn 44

Ethical, Legal, and Social Issues 45
Charles C. Carlson 45
Graham Chedd and Noel Schwerin 45
Debra L. Collins and R. Neil Schimke 45
Lane Conn 46
Jeff Davidson and Laurence Weinberger 47
Sharon Davis 47
Troy Duster 48
Rebecca S. Eisenberg 48
Stephen Goodman 49
Margaret C. Jefferson and Mary Ann Sesma 50
Mary B. Mahowald 50
Joseph D. McInerney and B. Ellen Friedman 51

Joseph D. McInerney, Lynda B. Micikas 52
Maureen M. Munn, Maynard V. Olson, and Leroy Hood 52
Philip J. Reilly, Dorothy C. Wertz, and Robin J.R. Blatt 53
Bari Scott 53
Maria Sosa 54
Sylvia J. Spengler 54
Sylvia Spengler and Janice Mann 55
Sara L. Tobin and Ann Boughton 55
Franklin M. Zweig 56
Infrastructure 59
Linda Holmes and Eugene Spejewski 59
Betty K. Mansfield and John S. Wassom 59
Sylvia J. Spengler 60
Walter Williams 61
James Wright 61
Small Business Innovation Research 63
Mark W. Knuth 63
Gualberto Ruano 63
Joseph Leone 64
William P. MacConnell 64
Ruth Ann Manning 64

Advanced Detectors for Mass Spectrometry

W.H. Benner and **J.M. Jaklevic**
Human Genome Group; Engineering Science Department; Lawrence Berkeley National Laboratory; University of California; Berkeley, CA 94720
510/486-7194, Fax: -5857, *whbenner@lbl.gov*
http://www-hgc.lbl.gov

Mass spectrometry is an instrumental method capable of producing rapid analyses with high mass accuracy. When applied to genome research, it is an attractive alternative to gel electrophoresis. At present, routine DNA analysis by mass spectrometry is seriously constrained to small DNA fragments. Contrasted to other mass spectrometry facilities in which the development of ladder sequencing is emphasized, we are exploring the application of mass spectrometry to procedures that identify short sequences. This approach helps the molecular biologists associated with LBL's Human Genome Center to identify redundant sequences and vector contamination in clones rapidly, thereby improving sequencing efficiency. We are also attempting to implement a rapid mass spectrometry-based screening procedure for PCR products.

The implementation of these applications requires that the performance of matrix-assisted-laser-desorption-ionization (MALDI) and electrospray mass spectrometry is improved. Our focus is the development of new ion detectors which will advance the state-of-the-art of each of these two types of spectrometers. One of the limitations for applying mass spectrometry to DNA analysis relates to the poor efficiency with which conventional electron multipliers detect large ions, a problem most apparent in MALDI-TOF-MS. To solve this problem, we are developing alternative detection schemes which rely on heat pulse detection. The kinetic energy of impacting ions is converted into heat when ions strike a detector and we are attempting to measure indirectly such heat pulses. We are developing a type of cryogenic detector called a superconducting tunnel junction device which responds to the phonons produced when ions strike the detector. This detector does not rely on the formation of secondary electrons. We have demonstrated this type of detector to be at least two orders of magnitude more sensitive, on an area-normalized basis, than microchannel plate ion detectors. This development could extend the upper mass limit of MALDI-TOF-MS and increase sensitivity.

Electrospray ion sources generate ions of mega-Dalton DNA with minimal fragmentation, but the mass spectrometric analyses of these large ions usually leads only to a mass-to-charge distribution. If ion charge was known, actual mass data could be determined. To address this problem, we are developing a detector that will simultaneously measure the charge and velocity of individual ions. We have been able to mass analyze DNA molecules in the 1 to 10 MDa range using charge-detection mass spectrometry. In this technique, individual electrospray ions are directed to fly through a metal tube which detects their image charge. Simultaneous measurement of their velocity provides a way to measure their mass when ions of known energy are sampled. Several thousand ions can be analyzed in a few minutes, thus generating statistically significant mass values regarding the ions in a sample population. We are attempting to apply this technology to the analysis of PCR products.

DOE Contract No. DE-AC03-76SF00098.

Mass Spectrometer for Human Genome Sequencing

Chung-Hsuan Chen, Steve L. Allman, and K. Bruce Jacobson
Oak Ridge National Laboratory; Oak Ridge, TN 37831
423/574-5895, Fax: -2115, *chenc@ornl.gov*

The objective of this program is to develop an innovative fast DNA sequencing technology for the Human Genome Project. It can also be applied to fast screening of genetic and contagious diseases, DNA fingerprinting, and environmental impact analysis.

The approach of this program is to replace conventional gel electrophoresis sequencing methods by using lasers and mass spectrometry for sequencing. The present gel sequencing method usually takes hours to days to acquire DNA analysis or sequencing, since different lengths of DNA segments need to be separated in dense gel. With laser desorption mass spectrometry (LDMS) approach, various sizes of DNA segments are separated in the vacuum chamber of a mass spectrometer. Thus, the time taken to separate various sizes of DNA is less than one second compared to hours using other methods.

Recently, we successfully demonstrated sequencing short DNA segments with this approach. We also have succeeded in using LDMS for fast screening of cystic fibrosis disease. We succeeded in identifying both point mutation and deletion of cystic fibrosis. In addition, we had preliminary success in using LDMS to achieve DNA fingerprinting. Thus, laser desorption mass spectrometry (LDMS) is going to emerge as a new and important biotechnological tool for DNA analysis.

DOE Contract No. DE-AC05-84OR21400.

***Projects designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).**

Genomic Sequence Comparisons

George Church
Harvard Medical School; Boston, MA 02115
617/432-0503 or -7562, Fax: -7266
http://arep.med.harvard.edu

The first objective of this project is completion of an automated system to sequence DNA using electrophore mass-tag (EMT) primers for dideoxy sequencing. The prototype machine will contain a 60 capillary array with 400 EMT-labeled sequence ladders per capillary. The system is designed to use 100-fold less reagent and have 500-fold higher speed (1000 bases per sec per instrument) than current sequencing technology. Cleavage and laser desorption of EMTs from membranes for subsequent detection by EC-TOF mass spectrometry. The second objective is to overcome the limitations of purely hypothetical annotation of the growing number of reading frames in new genome sequences. We measure gene product levels and interactions using DNA microarrays, whole genome in vivo footprinting and crosslinking.

Our approach involves system integration of instrumentation, organic chemistry, molecular biology, electrophoresis and software to the task of increasing sequencing accuracy and efficiency. Likewise we integrate such instruments and others with the needs of acquiring and annotation of large-scale microbial and human genomic sequence and population polymorphisms.

To establish functions for new genes, we use large scale phenotyping by multiplexed growth competition assays, both by targeted deletion and by saturation insertional mutagenesis. We will continue to develop a system to sequence DNA using electrophore mass-tags (EMTs). We will establish genome-scale experimental methods for sequence annotation.

The most significant findings in 1995-1996 were 1) Demonstration of use of electrophore mass-tags in dideoxy sequencing. 2) Development of IR-laser desorption method and model. 3) A novel dsDNA microarray synthesis strategy. 4) A new amplifiable differential display for whole-genome in vivo DNA-protein interactions. 5) Establishment and application of a microbial DNA-protein interaction database.

DOE Grant No. DE-FG02-87ER60565.

A PAC/BAC End-Sequence Data Resource for Sequencing the Human Genome: A 2-Year Pilot Study

Pieter de Jong
Roswell Park Cancer Institute; Buffalo, NY 14263
716/845-3168, Fax: -8849, *pieter@dejong.med.buffalo.edu*
http://bacpac.med.buffalo.edu

Large scale sequencing of the Human genome requires the availability of high-fidelity clones with large genomic inserts and a mechanism to find clones with minimal overlaps within the clone collections. The first need can be satisfied with bacterial artificial chromosome libraries (PACs and BACs) which already exist and further such libraries now being developed. However, a cost-effective way for establishing high-resolution contig maps for the human genome has not yet been established. Recently, a new approach for virtual screening for overlapping clones has been proposed by several research groups and has been discussed eloquently in a manuscript by Venter et al., 1996 (Nature). We will implement this approach for use with our human PAC and BAC libraries and use the first year as a pilot stage. The goal of the one year pilot is to prove the feasibility of large scale end sequencing and to demonstrate usefulness.

The first goal will be met by sequencing the ends for 40,000 clones from our existing PAC library and from BAC libraries currently being developed under NIH funding within our laboratory. The end-sequencing will be based on our new DOP-vector PCR procedure (Chen et al, 1996, Nucleic Acids Research 24, 2614-2616). All sequence data will be made available through public databases (GSDB, GDB, Genbank) and will also become BLAST searchable through the UTSW WWW site from our collaborator, Glen Evans. In view of our current under-developed informatics structure, we do not expect to provide BLAST search access through our own web site during the pilot phase.

To prove the usefulness of available end sequences, we will prepare a chromosome 14-enriched clone collection from our current 20-fold deep PAC library. To detect the chromosome 14 clones, we will use as hybridization probes a set of 1,000 mapped STS markers available from Paul Dear (MRC, Cambridge, UK), the about 600 markers present in the Whitehead map and the in situ mapped BAC and PAC clones available from Julie Korenberg. We will hybridize with these existing markers in probe pools, specific for regions of chromosome 14. Thus we will isolate region-enriched PAC clone collections.

Assuming that the clone collections will be at least 50%-specific for chromosome 14 (50% false positives) and will include most of the chromosome 14 PACs from our library, a collection of about 35,000 clones is expected.

Hence, the bulk of the end sequences obtained during the first year will be derived from the chromosome 14 enriched set and should result in a sequence ready clone collection covering about 100 Mbp of the human genome. The purity of the chromosome 14 PAC collection will be characterized in a number of different ways, including testing with independent markers not used as probes and by FISH analysis of a representative set of PAC clones. To test the usefulness of the end sequence resource, the Sanger Centre will sequence chromosome 14 PACs from our collection and identify overlapping clones by virtual screening, using our end-sequence database.

If overlapping clones can not be found with the expected level of redundancy in the end-sequence database, we will screen the original PAC library with probes or STS markers derived from the sequenced PAC clones.

Subcontract under Glen Evans' DOE Grant No. DE-FC03-96ER62294.

Multiple-Column Capillary Gel Electrophoresis

Norman Dovichi
Department of Chemistry; University of Alberta;
Edmonton, Alberta, Canada T6G 2G2
403/492-2845, Fax: -8231, *norm.dovichi@ualberta.ca*
http://hobbes.chem.ualberta.ca

The objective of this project is to develop high-throughput DNA sequencing instrumentation. A two-dimensional arrayed capillary electrophoresis instrument is under development.

We have developed multiple capillary DNA sequencers. These instruments have several important attributes. First, by operation at electric fields greater than 100 V/cm, we are able to separate DNA sequencing fragments rapidly and efficiently. Second, the separation is performed with 3%T 0%C polyacrylamide. This low viscosity, non-crosslinked matrix can be pumped from the capillary and replaced with fresh material when required. Third, we operate the capillary at elevated temperature. High temperature operation eliminates compressions, speeds the separation, and increases the read length. Fourth, our fluorescence detection cuvette is manufactured locally by means of microlithography technology. These detection cuvettes provide robust and precise alignment of the optical system. Currently, 5, 16, and 90 capillary instruments are in operation in our lab; 32 and 576 capillary devices are under development. Fourth, we use both avalanche photodiode photodetectors and CCD cameras for high sensitivity detection. We have obtained detection limits of 120 fluorescein molecules injected onto the capillaries. High sensitivity is important in detecting the low concentration fragments generated in long sequencing reads. This combination of low concentration acrylamide, high temperature operation, and high sensitivity detection allows separation of fragments over 800 bases in length in 90 minutes.

DOE Grant No. DE-FG02-91ER61123.

DNA Sequencing with Primer Libraries

John J. Dunn, Laura-Li Butler-Loffredo, and **F. William Studier**
Biology Department; Brookhaven National Laboratory;
Upton, NY 11973
516/344-3012, Fax: -3407, *dunn@genome1.bio.bnl.gov*
http://genome5.bio.bnl.gov

Primer walking using oligonucleotides selected from a library is an attractive strategy for large-scale DNA sequencing. Strings of three adjacent hexamers can prime DNA sequencing reactions specifically and efficiently when the template is saturated with a single stranded DNA-binding protein (1), and a library of all 4,096 hexamers is manageable. We would like to be able to sequence directly on 35-kbp fesmid templates, but the signal from a single round of synthesis is relatively weak and triple-hexamer priming has not yet been adapted for cycle sequencing. We reasoned that a hexamer library might be used for cycle sequencing if combinations of hexamers could be selectively ligated by using other hexamers as the template for alignment. In this way, the longer primers needed for cycle sequencing could be generated easily and economically without the need for complex machines for de novo synthesis.

We found that ordered ligation of 3 hexamers to form an 18-mer occurs readily on a template of the 3 complementary hexamers (offset by three base pairs) that can base pair unambiguously to form a double-stranded complex of indefinite length (2). Each hexamer forms three complementary base pairs with two other hexamers, generating complementary chains of contiguous hexamers with strand breaks staggered by three bases. Two adjacent hexamers in the chain to be ligated contain 5' phosphate groups and the others are unphosphorylated. Both T4 and T7 DNA ligase can ligate the phosphorylated hexamers to their neighbors in such a complex at hexamer concentrations in the 50-100 M range, producing an 18-mer and leaving three unphosphorylated hexamers. The products of these ligation reactions can be used directly for fluorescent cycle sequencing of 35-kbp templates.

Unambiguous ligation requires that alternative complexes with perfect base pairing not be possible with the combination of hexamers used. Since the combination of hexamers is dictated by the sequence of the desired ligation product, some oligonucleotides cannot be produced unambiguously by this method. However, 82.5% of all possible 18-mers could potentially be generated starting with a library of all

4096 hexamers, more than adequate for high throughput DNA sequencing by primer walking.

DOE Grant No. DE-AC02-76CH00016.

References

(1) Kieleczawa, J., Dunn, J. J., and Studier, F. W. DNA sequencing by primer walking with strings of contiguous hexamers. Science, 258, 1787-1791 (1992).

(2) Dunn, J. J., Butler-Loffredo, L. and Studier, F. W. Ligation of hexamers on hexamer templates to produce primers for cycle sequencing or the polymerase chain reaction. Anal.Biochem. 228, 91-100 (1995).

Rapid Preparation of DNA for Automated Sequencing

John J. Dunn, Matthew Randesi, and **F. William Studier**
Biology Department; Brookhaven National Laboratory; Upton, NY 11973
516/344-3012, Fax: -3407, *dunn@genome1.bio.bnl.gov*
http://genome5.bio.bnl.gov

We have developed a vector, referred to as a fesmid, for making libraries of approximately 35-kbp DNAs for mapping and sequencing. The high efficiency lambda packaging system is used to generate libraries of clones. These clones are propagated at very low copy number under control of the replication and partitioning functions of the F factor, which helps to stabilize potentially toxic clones. A P1 lytic replicon under control of the lac repressor allows amplification simply by adding IPTG. The cloned DNA fragment is flanked by packaging signals for bacteriophage T7, and infection with an appropriate T7 mutant packages the cloned sequence into T7 phage particles, leaving most of the vector sequence behind. The size of the vector portion is such that genomic fragments packageable in lambda (normal capacity 48.5 kbp) should also be packaged in T7 (normal capacity 40 kbp).

We have made fesmid libraries of several bacterial DNAs, including Borrelia burgdorferi (the cause of Lyme disease), Bartonella henselae (the cause of cat scratch fever), E. coli, B.subtilis, H. influenzae, and S. pneumoniae, some of which have been reported to be difficult to clone in cosmid vectors. Human DNA is also readily cloned in these vectors. Brief amplification followed by infection with a gene 3 and 17.5 double mutant of T7, which is defective in replicating its own DNA, produces lysates in which essentially all of the phage particles contain the cloned DNA fragment. Simple techniques yield high-quality DNA from these phage particles. Primers for direct sequencing from the ends of fesmid clones have been made.

Primer walking from the ends of fesmid clones could be an efficient way to sequence bacterial genomes, YACs, or other large DNAs without the need for prior mapping of clones. The ends of fesmids from a random library provide multiple sites to initiate primer walking. Merging of the elongating sequences from different clones will simultaneously generate the sequence of the original DNA and determine the order of the clones. The packaged fesmid DNAs are a convenient size for multiple restriction analyses to confirm the accuracy of the nucleotide sequence.

DOE Grant No. DE-AC02-76CH00016.

A PAC/BAC End-Sequence Database for Human Genomic Sequencing

Glen A. Evans, Dave Burbee, Chris Davies, Trey Fondon, Tammy Oliver, Terry Franklin, Lisa Hahner, Shane Probst, and Harold R. (Skip) Garner
Genome Science and Technology Center and McDermott Center for Human Growth and Development; University of Texas Southwestern Medical Center at Dallas; Dallas, TX 75235-8591
214/648-1660, Fax: -1666, *gevans@swmed.edu*
http://mcdermott.swmed.edu

While current plans call for completing the human genome sequence in 2003, major obstacles remain in achieving the speed and efficiency necessary to complete the task of mapping and sequencing. As an approach to this problem, we proposed a novel approach to large scale construction of sequence-ready physical clone maps of the human genome utilizing end-specific sequence sampling. An earlier pilot project was initially carried out to develop a GSS (genomic sequence sampled) map of human chromosome 11 by sequencing the ends of 17,952 chromosome 11 specific cosmids. This chromosome 11-specific end-sequence database allows rapid and sensitive detection of clone overlaps for chromosome 11-sequencing.

In this project, we propose to evaluate the utility of PAC and BAC end-sequences representing the entire human genome as a tool for complete, high accuracy mapping and sequencing. In this approach, we utilized total genomic PAC/BAC libraries (constructed by P. de Jong, RPCI), followed by end-sequencing of both ends of each clone in the library and limited regional mapping of a subset of clones as sequencing nucleation points by FISH (Fluorescence in situ hybridization).

To initiate regional analysis, a single clone would be sequenced by shotgun or primer directed sequencing, the entire sequence used to search the end-database for overlapping clones, and the minimal overlapping clones for extending the sequence selected. This approach would allow rational and efficient simultaneous mapping and sequencing, as well as expediting the coordination and exchange of information between large and small groups participating in the human genome project.

In this pilot project proposal we are carrying out automated end-sequencing of approximately 40,000 PAC and BAC clones representing the entire human genome, as well as about 500 PAC clones localized to human chromosomes 11 and 15. The clones and resulting end-sequence data base will be utilized to 1) nucleate regions of interest for large scale sequencing concentrating on regions of chromosome 11 and 15, 2) correspond with regions mapped by other methods to confirm the mapping accuracy and 3) used to evaluate the use of random clone end sequence libraries. DNA sequencing is being carried out in an entirely automated fashion using a Beckman/Sagian robotic system, ABI 377 automated sequencers and automated sequence data processing, annotation and publication using a Hewlett Packard/Convex superparallel computer located at the UTSW genome center. FISH analysis of a sample of PAC clones has been carried out and defines the potential chimera rate in existing PAC libraries as less than 1.2%. This effort will be coordinated with efforts of other groups carrying out PAC and BAC library construction, PAC and BAC end-sequencing and FISH analysis to avoid duplication of effort and provide a comprehensive end-sequence library and data set for use by the international human genome sequencing effort.

DOE Grant No. DE-FC03-96ER62294.

Automated DNA Sequencing by Parallel Primer Walking

Glen A. Evans, Dave Burbee, Chris Davies, Jeff Schageman, Shane Probst, Terry Franklin, Ken Kupfer, and Harold R. (Skip) Garner
Genome Science and Technology Center and McDermott Center for Human Growth and Development; University of Texas Southwestern Medical Center at Dallas; Dallas, TX 75235-8591
214/648-1660, Fax: -1666, *gevans@swmed.edu*
http://mcdermott.swmed.edu

The development of efficient mapping approaches coupled with high throughput, automated DNA sequencing remains one of the key challenges of the Human Genome Project. Over the past few years, a number of strategies to expedite clone-by-clone DNA sequencing have been developed including efficient shotgun sequencing, sequencing of nested deletions, and transposon-mediated primer insertion. We have developed a novel sequencing strategy applicable to high throughput, large scale genomic analysis based upon DNA sequencing directly primed on of cosmid templates using custom-designed, automatically synthesized oligonucleotide primers. This approach of directed primer "walking" would allow the number of sequencing reactions and the efficiency of sequencing to be vastly improved over traditional shotgun sequencing.

Custom primer design has been carried out using software we developed for prediction of "walking" primers directly from the output of ABI377 automated DNA sequencers, and the output used to automatically program synthesis of the custom primers using 96 or 192 channel oligonucleotide synthesizers constructed at UTSW. Automated operation of the sequencing system is thus possible where results of each sequencing reaction is used to predict, synthesize, and carry out appropriate extension reactions for downstream "walking". A automated prototype system has been assembled where dye terminator DNA sequencing can be carried out from 96 cosmid templates simultaneously followed by prediction of oligonucleotide "walking" primers for extending the sequence of each fragment, and programming an attached 96-channel oligonucleotide synthesizer to initiate a second round of sequencing. Using a set of nested cosmids covering 800 kb at 5X redundancy, primer directed sequencing should allow completion of 800 kb of finished, high accuracy DNA sequence in 8 to 16 cycles. Furthermore, coupling of automated DNA sequencing instrumentation to DNA sequence analysis programs and multichannel oligonucleotide synthesizers will allow almost complete automation of sequencing process and the development of instrumentation for completely unattended DNA sequencing.

DOE Grant No. DE-FG03-95ER62055.

*Parallel Triplex Formation as Possible Approach for Suppression of DNA-Viruses Reproduction

V.L. Florentiev, A.K. Shchyolkina, I.A. Il'icheva, E.N. Timofeev, and S. Yu Tsybenko
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
Fax: +7-095/135-1405, *flor@imb.imb.ac.ru*

It is well known that homopurine or homopyrimidine single stranded oligonucleotides can bind to homopurine-homopyrimidine sequences of two-stranded DNA to form stable three-stranded helices. In such triplexes two identical strands have antiparallel orientation. We denote these triplexes as "antiparallel" or "classical" triplexes.

A particular interest of investigators to triplexes has arisen due to an elegant idea of using triplexes as sequence-specific tools for purposeful influence on DNA duplexes. Triplex forming oligonucleotides were shown to be potentially useful as regulators of gene expression and subsequently as therapeutical (antiviral) agents.

A significant limitation to the practical application of antiparallel triplex is the requirement for homopurine tracts in target DNA sequences. Numerous investigations slightly expanded the repertoire of triple-forming sequences but did not completely remove this limitation.

It was recently shown that during homologous recombination promoted by RecA a triple-stranded DNA intermediate was formed. Such a structure is a new form of the triple helix. In sharp contrast with the "classical" triplexes their third strand is parallel to the identical strand of the Watson-Crick duplex. We denote this structure as "parallel" triplex. Recently, the parallel triplex was obtained only by deproteinization of joint molecules generated by recombination proteins.

We first obtained experimental (chemical probe, melting curves and fluorescence due binding) results that provide convincingly evidence for protein-independent formation of parallel triplex [1] and than confirmed this fact by FTIR data [2]. Because the parallel triplex can be formed for any sequence, it might be "ideal" potential tool for sequence specific recognition of DNA. Unfortunately, low stability of parallel triplexes prohibits practical application of these structures.

Earlier we found that propidium iodide stabilizes selectively the parallel triplexes [3]. This fact was the basis of new approach to stabilization of parallel triplexes being developed by us now. The approach consists in use of targeting oligonucleotide, which contains in internucleotide linkage the alkyl insert coupled with intercalated ligand through linker. Length of linker was chosen to allow ligand to intercalate in the same stacking-contact (length of linker was picked by molecular dynamic calculations).

Preliminary study showed that presence of intercalating inserts increase considerably stability of DNA duplexes [4]. Now we are investigating in detail effect of such modification of targeting oligonucleotides on stability of parallel triplexes.

DOE Grant No. OR00033-93CIS005.

References

1. Shchyolkina, A. K., Timofeev, E. N., Borisova, O. F., Il'icheva, I.A., Minyat, E. E., Khomyakova, E. B. and Florentiev, V. L. (1994) The R-form DNA does exist. FEBS Letters, 339, 113-118.
2. Dagneaux, C., Gousset, H., Shchyolkina, A. K., Ouali, M., Lettelier, R., Liquier, J., Florentiev, V. L. and Taillander, E. (1996) Parallel and antiparallel AA-T intramolecular triple helices. Nucleic Acids Res., 24, 4506-4512.
3. Borisova, O. F., Shchyolkina, A. K., Timofeev, E. N., Tsybenko, S. Yu., Mirzabekov, A. and Florentiev, V. L. (1995) Stabilization of parallel triplex with propidium iodide. J. Biomol. Struct. Dynam., 13, 15-27.
4. Timofeev, E. N., Smirnov I. P., Haff, L. A., Tishchenko, E. I., Mirzabekov, A. D. and Florentiev, V. L. (1996) Methidium intercalator inserted into synthetic oligonucleotides. Tetrahedron Lett., 37, 8467-8470.

Advanced Automated Sequencing Technology: Fluorescent Detection for Multiplex DNA Sequencing

Andy Marks, Tony Schurtz, F. Mark Ferguson, Leonard Di Sera, Alvin Kimball, Diane Dunn, Doug Adamson, Peter Cartwright, Robert B. Weiss,[1] and **Raymond F. Gesteland**[1]
Department of Human Genetics and [1]Howard Hughes Medical Institute; University of Utah; Salt Lake City, UT 84112
Gesteland: 801/581-5190, Fax: /585-3910
ray.gesteland@genetics.utah.edu

Automation of a large-scale sequencing process based on instrumentation for automated DNA hybridization and detection is a focal point of our research. Recently, we have devised a method for amplifying fluorescent light output on nylon membranes by using an alkaline phosphatase-conjugated probe system combined with a fluorogenic alkaline phosphatase substrate [1]. The amplified signal allows sensitive detection of DNA hybrids in the sub-femtomole/band range.

On the basis of this detection chemistry, automated devices for detecting DNA on blotted microporous membranes using enzyme-linked fluorescence, termed Probe Chambers, have been built. The fluorescent signal is collected by a CCD camera operating in a Time Delay and Integration mode. Concentrated solutions of probes and enzymes are stored in Peltier-cooled septa sealed vials and delivered by syringe pumps residing in a gantry style pipetting robot. Fluorescence excitation is generated by a mercury arc lamp acting through a fiber optic "light line". Three 30 x 63 centimeter sequencing membranes can be simultaneously processed, currently revealing up to 108 lane sets per multiplex cycle. A probing cycle is completed approximately every eight hours.

Integration of the Probe Chamber into the production pipe line is accomplished through connections to the laboratory data base. A critical component of a high-throughput sequencing laboratory is the software for interfacing to instrumentation and managing work flow. The Informatics Group of the Utah Genome Center has designed and implemented an innovative system for automating and managing laboratory processes. This software allows the model of workflow to be easily defined. Given such a model, the system allows the user to direct and track the flow of laboratory information. The core of the system is a generic, client-server process management engine that allows users to define new processes without the need for custom programming. Based on these definitions, the software will then route information to the next process, track the progress of each task, perform any automated operations, and provide reports on these processes. To further increase the usefulness of our laboratory information system, we have augmented it with hand-help mobile computing devices (Apple Newtons) that link to the database through RF networking cards.

Base calling software has been developed to support our automated, large scale sequencing effort. 1st stage sequence calling identifies putative bands, however, depending on the number of reader indel errors (2-6%), merging 1st stage sequence without the aide of cutoff information can be difficult. To improve our base calling we have employed Fuzzy Logic to establish confidence metrics. The logic produces a confidence metric for each band using band height, width, uniqueness, shape, and the gaps to adjacent bands. The confidence metric is then used to identify the largest block of highest quality sequence to be merged.

DOE Grant No. DE-FG03-94ER61817.

Reference

[1] Cherry, J.L., Young, H., Di Sera, L.J., Ferguson, F.M., Kimball, A.W., Dunn, D.M., Gesteland, R.F., and Weiss, R.B. (1994). Enzyme-linked fluorescent detection for automated multiplex DNA sequencing, Genomics 20, 68-74.

Resource for Molecular Cytogenetics

Donna Albertson, Colin Collins, **Joe Gray**,[1] Steven Lockett, **Daniel Pinkel**,[1] Damir Sudar, Heinz-Ulrich Weier, and Manfred Zorn
Lawrence Berkeley National Laboratory; Berkeley, CA 94720 and [1]University of California; San Francisco, CA 94143
Gray: 415/476-3461, Fax: -8218, *gray@cc.ucsf.edu*
Pinkel: 415/476-3659, Fax: -8218, *pinkel@cc.uscf.edu*
http://rmc-www.lbl.gov

The purpose of the Resource for Molecular Cytogenetics is to develop molecular cytogenetic techniques, instruments and reagents needed to facilitate large scale genomic DNA sequencing and to assist in identification and functional characterization of genes involved in disease susceptibility, genesis and progression. This work is closely coordinated with the LBNL Human Genome Program and directly supports research in the LBNL Life Sciences Division and the UCSF Cancer Center. Work currently is in four areas: a)Genome analysis technology, b)Probe development and physical map assembly, c)Digital imaging microscopy and d)Informatics. The Resource acts as a catalyst for research in several areas so some support comes from Industry, the NIH and NIST.

Probe development and physical map assembly: The Resource maintains a list of over a thousand publicly available probes suitable for molecular cytogenetic studies. These include approximately 600 probes each selected by the Resource to contain a known STS or EST. Probes selected by the Resource can be requested through our web page. The Resource also participates in the development of low and high resolution physical maps to facilitate analysis and characterization of genetic abnormalities associated with human disease. Low resolution mapping panels with probes distributed at few megabase intervals have been completed this year for chromosomes 1, 2, 3, 7, 8, 10, and 20. The mapped STSs associated with these probes facilitate movement from low to high resolution physical maps. STS content mapping and DNA fingerprinting have been applied to develop a high resolution, sequence-ready map comprised of BAC and P1 clones for the ~1Mb region of chromosome 20 between WI9227 and D20S902. This region is amplified in ~10% of human breast cancers. Approximately 300 kb of this region has been sequenced by the LBNL Human Genome Program.

Quantitative DNA fiber mapping (QDFM) has been developed this year to facilitate high resolution analysis of genomic overlap between cloned probes. In this approach, cloned DNA molecules are uniformly stretched during drying by the hydrodynamic action of a receding meniscus. The position of specific sequences along the stretched DNA molecules is visualized by fluorescence in situ hybridization (FISH) and measured by digital image analysis. QDFM has been used to map gamma alpha transposons, plasmid or cosmid probes along P1 molecules, and P1 or PAC clones along straightened YAC molecules with few kilobase resolution. QDFM is now being studied to determine its utility in the assembly of minimally overlapping, sequence-ready contigs, assessment of the integrity of cloned BACs and mapping of subclones prepared for directed DNA sequencing along the clone from which they were derived.

Genome analysis technology: The Resource has participated in the development of comparative genomic hybridization (CGH) as a tool for detection and mapping of changes in relative DNA sequence copy number in humans and mouse. This year, CGH to arrays of cloned probes (CGHa) has been demonstrated. This is advantageous because it allow aberrations to be mapped with resolution determined by the genomic spacing of probes on the array. CGHa also is attractive since it appears to be linear over a relative copy number range of at least 104 between the two nucleic acid samples being compared.

The Resource has participated in the development of FISH approaches to analysis of relative gene expression in normal and aberrant tissues. FISH with cloned or predicted expressed sequences, previously developed in C. elegans, is now being applied to the assessment of expression of human genes. The C. elegans work suggests a throughput of several dozen sequences per month. Information from this approach will be important in assessment of the function of newly discovered genes, including those predicted from DNA sequencing.

(abstract continued)

Digital imaging microscopy: The Resource supports work in microscopy, image processing and analysis methods needed for CGH and CGHa, 3D FISH, tissue analysis, rare event detection, multi-color image acquisition, aberration scoring for biodosimetry, and analysis of FISH to DNA fibers. Developments this year include an improved package for CGH and prototype systems for analysis of DNA fibers, CGHa arrays and semiautomatic segmentation of nuclei in three dimensions.

Informatics: The Resource maintains a web site at http://rmc-www.lbl.gov that summarizes information about mapped probes. Probes developed by the Resource can be requested directly through this page. In addition, the Resource has developed a Web page for exchange of genomic, genetic and biologic information between geographically disperse collaborators. The page, under password control, carries information about physical maps, genomic sequence, sequence annotation, and gene expression images.

DOE Contract No. DEAC0376SF00098.

DNA Sample Manipulation and Automation

Trevor Hawkins
Center for Genome Research; Whitehead Institute/Massachusetts Institute of Technology; Cambridge, MA 02139
617/252-1910, Fax: -1902, *tlh@genome.wi.mit.edu*
http://www-genome.wi.mit.edu

The objective of this project is to develop a high-throughput, fully automated robotic device for the complete automation of the sequencing process. We also aim to further develop DNA sequencing electrophoresis systems and to integrate these devices with our robotics.

We have built the Sequatron, an integrated, robotic device which automates the tasks of DNA purification and setup of thermal cycle sequencing reactions. The major component of our system is an articulated CRS 255A robotic arm which is track mounted. The deck of the robot contains several new or modified XYZ robotic workstations, a novel thermal cycler with automated headed lids, carousels, and custom built plate feeders.

Biochemically, we have employed our Solid-phase reversible immobilization (SPRI) technique to isolate and manipulate the DNA throughout the process.

Specifically we have set up the Sequatron to isolate DNA from M13 phage or crude PCR products using the same protocol and procedures. From M13 phage we obtain approximately 1g of DNA per well, which is sufficient for multiple sequencing reactions.

The current throughput of the system is 80 microtiter plates of samples from M13 phage supernatants or crude PCR products to sequence ready samples every 24 hours. Recently, new enzymes, new energy transfer primers and higher density microtiter plates have opened up possible increases to in excess of 25,000 samples per 24 hour period.

DOE Grant No. DE-FG02-95ER62099.

Relevant Publication

DeAngelis, M., Wang, D., & Hawkins, T. (1995) Nucl. Acids. Res 23, 4742-4743.

Construction of a Genome-Wide Characterized Clone Resource for Genome Sequencing

Leroy Hood, **Mark D. Adams**,[1] and **Melvin Simon**[2]
University of Washington; Seattle, WA 98195-7730
206/616-5014, Fax: /685-7301, *tawny@u.washington.edu*
[1]The Institute for Genomic Research; Rockville, MD 20850; *mdadams@tigr.org*
[2]California Institute of Technology; Pasadena, CA 91125; *simonm@starbase1.caltech.edu*

Bacterial artificial chromosomes (BACs) represent the state of the art cloning system for human DNA because of their stability and ease of manipulation. Venter, Smith and Hood (Nature 381:364-366, 1996) have proposed a strategy based on the use of sequences from the ends of all clones in a deep coverage BAC library to produce a sequence-ready set of clones for the human genome. We propose to demonstrate the effectiveness of this strategy by performing a directed test, initially on chromosomes 16 and 22, and continuing on to chromosome 1. All available markers on chromosome 16 (including the large number of soon-to-be-available radiation hybrid markers) will be used to screen the existing 8x BAC library at CalTech. This will serve to evaluate the quality of the library in terms of representation of broad chromosomal regions. A similar procedure will be used for chromosome 22, except that the existing BAC map will be used to select more evenly spaced markers for screening, including use of end-sequence markers from the current chromosome 22 BAC map constructed in the Simon lab. Each identified clone will be rearrayed from the library and end sequenced. This information will dovetail nicely with ongoing sequencing projects at TIGR and the Sanger Centre, which will in turn provide additional information on the average degree of BAC overlap detectable by this method, the degree of interference with genome-wide repeats, and the appropriate use of fingerprinting as an early or late addition to the end-sequencing information. In addition, we will develop and implement cost-effective, high-throughput methods of preparing and end-sequencing BAC DNA that are suitable for scaling to characterization of the full 400,000 clones necessary for characterization of a 15x human BAC library.

DOE Grant No. DE-FC03-96ER62299.

DNA Sequencing Using Capillary Electrophoresis

Barry L. Karger
Barnett Institute; Northeastern University; Boston, MA 02115
617/373-2867 or -2868, Fax: -2855
bakarger@lynx.neu.edu

During the past year, we have made major progress in the design of a replaceable polymer matrix for DNA sequencing and the development of the first generation multiple capillary array of 12 capillaries. We also implemented ultrafast separation of dsDNA (e.g. 30 sec for complete resolution of the standard X174-HAE III restriction fragments).

In the separation of sequencing reaction products, we completed a study on the role of polymer molecular weight and concentration. Using linear polyacrylamide (LPA), the polymer with which we have had our most success, we have achieved 1000 base read lengths in 1 1/2 hrs. Optimization of column length, electric field and column temperature (50° C) was required. Using emulsion polymerization, we are now able to produce LPA powders with MW of $\sim 10^4$ k Da. The fully replaceable matrix is very powerful for rapid sequencing of long reads.

We have successfully implemented a 12-capillary array instrument and are using it to study issues of ruggedness in routine sequencing. As part of this, we have developed a sample clean-up procedure which reduces all reactions to a similar state in terms of sample solution prior to injection. The results of this work have led to the design of a 96-capillary array that we will implement over the next year.

We have also achieved very fast separations of ss- and dsDNA using short capillaries and very high yields. For example, sequencing 300 bases in 3–4 mins. has been shown, as well as very rapid mutational analysis. Implementation of such speeds on a capillary array will create an instrument for high throughput automated analysis.

DOE Grant No. DE-FG02-90ER60985.

Ultrasensitive Fluorescence Detection of DNA

Richard A. Mathies and **Alexander N. Glazer**
Departments of Chemistry and Molecular and Cell Biology; University of California; Berkeley CA 94720
510/642-4192, Fax: -3599, *rich@zinc.cchem.berkeley.edu*

The overall goal of this project is to develop new fluorescence labeling methods, separation methods and detection technologies for DNA sequencing and genomic analysis.

Highlights along with representative publications are given below.

Energy Transfer Primers. Families of sequencing and PCR primers have been developed that contain both fluorescence donor and acceptor chromophores.[1] These labeled primers with optimized excitation and emission properties provide from 2- to 20-fold enhanced signal intensities in automated DNA sequencing with slab gels and with capillary arrays.[2] The reduced spectral cross talk of these ET primers also makes them valuable in PCR product and STR analyses.[3]

New Intercalation Dye Labels. A new family of heterodimeric bis-intercalation dyes has been synthesized exploiting the concept of fluorescence energy transfer between two different cyanine intercalators.[4] By tailoring the spectroscopic properties of the dyes, labels with intense emission above 650 nm following 488 nm excitation have been fabricated. By adjusting the spacing linker between the two dyes, the binding affinity has also been optimized. These molecules are useful for noncovalent multiplex labeling of ds-DNA in a wide variety of multicolor analyses.[5]

Capillary Electrophoresis Chips. Capillary and capillary array electrophoresis systems have been photolithographically fabricated on 2x3' glass substrates.[6] These devices provide high quality electrophoretic separations of ds-DNA fragments and DNA sequencing reactions with a 10-fold increase in speed.[7] Arrays of up to 32 capillaries on a single chip have been fabricated.

Single DNA Molecule Fluorescence Burst Detection. A confocal fluorescence system has been used to demonstrate that single molecule fluorescence burst counting can be used to detect CE separations of ds-DNA fragments. Fragments as small as 50 bp can be counted and mass sensitivities as low as 100 molecules per electrophoresis band are possible. This technology should be valuable in incipient cancer and trace pathogen detection.[8]

DOE Grant No. DE-FG03-91ER61125.

(abstract continued)

References

1. Ju, J., Ruan, C., Fuller, C. W., Glazer, A. N. and Mathies, R. A. Fluorescence Energy Transfer Dye-Labeled Primers for DNA Sequencing and Analysis, Proc. Natl. Acad. Sci. U.S.A. 92, 4347-4351 (1995).
2. Ju, J., Glazer, A. N. and Mathies, R. A. Energy Transfer Primers: A New Fluorescence Labeling Paradigm for DNA Sequencing and Analysis, Nature Medicine 2, 180-182 (1996).
3. Wang, Y., Ju, J., Carpenter, B., Atherton, J. M., Sensabaugh, G. F. and Mathies, R. A. High-Speed, High-Throughput THO1 Allelic Sizing Using Energy Transfer Fluorescent Primers and Capillary Array Electrophoresis, Analytical Chemistry 67, 1197-1203 (1995).
4. Benson, S. C., Zeng, Z., and Glazer, A. N. Fluorescence Energy Transfer Cyanine Heterodimers with High Affinity for Double-Stranded DNA. I. Synthesis and Spectroscopic Properties, Anal. Biochem. 231, 247-255 (1995).
5. Zeng, Z., Benson, S. C., and Glazer, A. N. Fluorescence Energy Transfer Cyanine Heterodimers with High Affinity for Double-Stranded DNA. II. Applications to Multiplex Restriction Fragment Sizing, Anal. Biochem. 231, 256-260 (1995).
6. Woolley, A. T. and Mathies, R. A. Ultra-High-Speed DNA Fragment Separations Using Microfabricated Capillary Array Electrophoresis Chips, Proc. Natl. Acad. Sci. U.S.A., 91, 11348-11352 (1994).
7. Woolley, A. T. and Mathies, R. A. Ultra-High-Speed DNA Sequencing Using Capillary Array Electrophoresis Chips, Analytical Chemistry 67, 3676-3680 (1995).
8. Haab, B. B. and Mathies, R. A. Single Molecule Fluorescence Burst Detection of DNA Fragments Separated by Capillary Electrophoresis, Analytical Chemistry 67, 3253-3260 (1995).

Joint Human Genome Program Between Argonne National Laboratory and the Engelhardt Institute of Molecular Biology

Andrei Mirzabekov,[1,2] G. Yershov,[1,2] Y. Lysov,[2] V. Barsky,[2] V. Shick,[2] and S. Bavikin[1]
[1]Argonne National Laboratory; Argonne, Il 60439
630/252-3161 or -3361, Fax: /252-3387
amir@everest.bim.anl.gov
[2]Engelhardt Institute of Molecular Biology; 117984 Moscow, Russia

In 1996, more than thirty U.S. and Russian research workers participated in the joint Human Genome Program between Argonne National Laboratory and Engelhardt Institute of Molecular Biology on the development of sequencing by hybridization with oligonucleotide microchips (SHOM).

During this year, about twenty Russian scientists have been working from 3 months to 1 year in ANL. In this period, 3 papers have been published and 5 papers accepted for publication, 3 more papers are submitted for publication.

The main research efforts of the group have been concentrated in three directions:

I. Improvement of SHOM technology.
II. Development of SHOM for the needs of Human Genome Program.
III. Development of new approaches based on SHOM technology.

I. Improvement of SHOM technology

As a major result of the work in this direction, simple, reliable and effective methods of microchip manufacturing, sample preparations, and quantitative hybridization analysis by fluorescence microscopy have been developed or improved.

1. Photopolymerization technique for production of micromatrices of polyacrylamide gel pads on hydrophobicized glass surface was improved to become a simple, highly reproducible and inexpensive procedure (7).

2. New and cheaper chemistry of the oligonucleotide immobilization has been developed and introduced for production of more durable microchips. It is based on the use of amino-oligonucleotides and aldehyde-gels instead of 3-methyluridine-oligonucleotides and hydrazide-gels (3).

3. Four-pin robot has been constructed with computer control of every microchip element production. High quality microchips with 4100 immobilized oligonucleotides have been manufactured and the complexity of the microchips can easily be scaled up to a few tens of thousand elements.

4. Two-color fluorescence microscope has been equipped for regular use with proper mechanics and software. It allows investigators to regularly use the automatic quantitative monitoring of the hybridization on the whole microchip and to measure the kinetics of hybridization as well as the melting curves of duplexes formed with all microchip oligonucleotides (1,2,8).

5. Four-color fluorescence microscope was manufactured and four proper fluorescence dyes are at present under selection.

6. Chemical methods of introduction of several fluorescence dyes into DNA and RNA with or without fragmentation have been developed and regularly used in SHOM experiments (4).

7. A theory describing the kinetics of hybridization with gel-immobilized oligonucleotides has been developed (5).

8. Simple and relatively inexpensive equipment (around $10,000 per set) has been produced for manual manufacturing of microchips and fluorescence measurement of hybridization, which will enable every laboratory to produce and practically use microchips containing up to 100 immobilized oligonucleotides or other compounds.

II. Application of SHOM

Although the main goal of our SHOM development is to produce a simple de novo sequencing procedure, a number of other SHOM applications have been tested as intermediate steps in the SHOM research.

1. Sequence analysis and sequencing

A number of technical problems should be solved for de novo sequencing although they are much less stringent for comparative sequence analysis than for de novo sequencing. Among these:

a) Reliable discrimination of perfect and mismatched duplexes. We have significantly improved the discrimination by decreasing the length of hybridized oligonucleotides to 6-and 8-mers (1, 7) and by using 5-mers in "contiguous stacking" hybridization (1,2). Essential improvement was also achieved by automatic measuring of the melting curves for duplexes formed in each microchip element and calculating their thermodynamic parameters, free energy, enthalpy and entrophy for different regions of the melting curves and by comparing them with these parameters for perfect duplexes. In addition, a highly reliable discrimination was achieved by using two-color fluorescence microscopy and by quantitative comparison of the hybridization pattern of a known DNA or synthetic oligonucleotides and DNA under study labeled with different fluorophores (8).

b) Difference in hybridization efficiency depends on the GC-content and the length of the duplex. We have equalized the efficiency by choosing proper concentration for the immobilized oligonucleotide (6,7) and also by increasing the effective length of immobilized oligonucleotides by adding at one or both their ends 5-nitroindole as a universal base or a mixture of four bases (2).

c) Interference of hairpins and other structures in DNA with less stable duplexes formed upon the DNA hybridization with comparatively short immobilized oligonucleotides of the microchip. This interference was decreased by fragmentation of the analysed sample of DNA and RNA in the course of incorporation of a fluorescence label (4). We have also tested incorporation by a chemical bond of an intercalator into immobilized oligonucleotides that stabilized its base paring with DNA over hairpin formation (10).

d) Necessity to increase the microchip complexity for sequencing long DNA stretches. As an alternative, further development of so-called contiguous stacking hybridization was shown to improve the efficiency of 8-mer microchip up to that of 13-mer microchip so that DNA of several kilobases in length could be sequenced by SHOM (2).

e) 6-mer microchips for sequencing and sequence analysis. We have now come to the stage of manufacturing microchips containing 4,096 (i.e. all possible) 6-mers. The control tests partly described above have shown that these microchips can be effectively used for sequence analysis, mutation diagnostics and detection of sequencing mistakes by conventional gel-sequencing methods. We hope that after demonstrating the efficiency of 6-mer microchips, we shall be able to get sufficient financial support for production of the microchip with all 65,536 8-mers.

2. Mutation diagnostics and gene polymorphism analysis

The improvements described above have been introduced for reliable ("Yes" or "No" mode) identification of single-base changes in human genomic DNA. The efficiency of SHOM has been demonstrated for identification of a number of b-thalassemia mutations (1,2,8) and HLA allele variations in the human genome.

3. Identification of microorganisms and gene expression monitoring

Bacterial microchips have been manufactured and tested. Their ability for reliable identification of a number of bacterial strains in the sample has been demonstrated (6). The chips containing oligonucleotides complementary to specific regions of 16S ribosomal RNA were hybridized with samples of rRNA, total RNA, DNA and RNA transcripts of PCR-amplified genomic rDNA. Similar preliminary experiments demonstrated the efficiency of SHOM for monitoring the gene expression.

III. Development of new approaches based on the SHOM technology

1. Enzymatic modification of nucleic acids on selected elements of the oligonucleotide chip. The gel pads of the oligonucleotide chip are separated from each other by hydrophobic glass surface. It prevents the cross-talking of the chip elements when a drop of solution is applied on specified elements. At the same time, a high porosity of the gel allows diffusion of large proteins into the gel. We have demonstrated that immobilized oligonucleotides can be enzymatically phosphorylated and ligated with contiguously stacked 5-mer after hybridization with DNA. A walking sequencing procedure by stacked pentanucleotides was proposed that is based on enzymatic ligation and phosphorylation on oligonucleotides chips (9).

2. DNA fractionation on oligonucleotide chips. Due to the same properties, the oligonucleotide chips are used for fractionation of DNA after DNA hybridization with some complementary oligonucleotides of the chip. A new procedure for sequencing long DNA pieces was proposed that is based on fractionation of DNA on fractionating oligonucleotide chips followed by sequencing of the isolated DNA by SHOM on sequencing microchips. The procedure allows the investigator to skip cloning and mapping of long DNA pieces (9).

Conclusions

It appears that the major technical problems of SHOM have been in most part solved, and this technology can already be applied for sequence analysis and checking the accuracy of conventional sequencing methods. A number of other applications in the Human Genome Program are within the reach of SHOM, such as mutation screening, gene polymorphism studies, detection of microorganisms, gene expression studies, etc. Application of SHOM for de novo DNA sequencing requires manufacturing of more complicated microchips and improvement of some other, already available methods.

DOE Contract No. W-31-109-Eng-38.

References

1. Yershov G., Barsky V., Belgovsky A., Kirillov Eu., Kreindlin E., Ivanov I., Parinov S., Guschin D., Drobishev A., Dubiley S., Mirzabekov A. DNA analysis and diagnostics on oligonucleotide microchips. // Proc. Natl. Acad. Sci. 1996. Vol. 93. 4913-4918.
2. Parinov S., Barsky V., Yershov G., Kirillov Eu., Timofeev E., Belgovskiy A., Mirzabekov A. DNA sequencing by hybridization to microchip octa-and decanucleotides extended by stacked pentanucleotides. // Nucl. Acids Res. 1996. Vol. 24. N 15. P. 2998-3004.
3. Timofeev E., Kochetkova S., A., Mirzabekov A. Radioselective immobilization of short oligonucleotides to acrylic copolymer gels // Nucl. Acids Res. 1996. Vol. 24. N 16. P. 3142-3148.
4. Prudnikov D., Mirzabekov A. Chemical methods of DNA and RNA fluorescent labelling. // Nucl. Acids Res. 1996., in press.
5. Livshits M., Mirzabekov A. Theoretical analysis of the kinetics of DNA hybridization with gel-immobilized oligonucleotides. // Biophys. J. 1996. Vol. 71, in print//.
6. Guschin D., Mobarry B., Proudnikov D., Stahl D., Rittmann B., Mirzabekov A. Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology //Applied and Environmental Microbiology, in print//.
7. Guschin D., Yershov G., Zaslavsky A., Gemmell A., Shick V., Lysov Yu., Mirzabekov A. A simple method of oligonucleotide microchip manufacturing and properties of the microchips // submitted for publication.
8. Drobyshev A., Mologina N., Shik V., Pobedimskaya D., Yershov G., Mirzabekov A. Sequence analysis by hybridization with oligonucleotide microchip: identification of beta-thalassemia mutations // Gene (in print).
9. Dubiley S., Kirillov Eu., Lysov Yu., Mirzabekov A. DNA fractionation, sequence analysis and ligation of immobilized oligomers on oligonucleotide chips // submitted for publication.
10. Timofeev E., Smirnov I.P., Haff L.A., Tishchenko E.I., Mirzabekov A.D., Florentiev V.L.. Methidium Intercalator Inserted into Synthetic Oligonucleotides // Tetrahedron Letters 1996, v.37, N47, p.8467.

Relevant Publication

Methods of DNA sequencing by hybridization based on optimizing concentration of matrix-bound oligonucleotide and device for carrying out same by Khrapko K., Khorlin A., Ivanov I., Ershov G., Lysov Yu., Florentiev V., Mirzabekov A. US Patent 5,552,270, Sep. 3, 1996. PCT/RU92/00052, filed Mar 18, 1992.

High-Throughput DNA Sequencing: SAmple SEquencing (SASE) Analysis as a Framework for Identifying Genes and Complete Large-Scale Genomic Sequencing

Robert K. Moyzis and Jeffrey K. Griffith[1]
Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3912, Fax: -2891, *moyzis@telomere.lanl.gov*
[1]University of New Mexico; Albuquerque, NM 87131

The human chromosome 5 and 16 physical maps (Doggett et al., Nature 377:Suppl:335-365, 1995; Grady et al., Genomics 32:91-96, 1996) provide the ideal framework for initiating large-scale DNA sequencing. These physical mapping studies have shown clearly that gene density in humans will vary greatly. For example, band 16q21, consisting of 8 Mb of DNA, has no genes or trapped exons assigned to it, as yet. In contrast, band 16p13.3 has an extremely high density of coding regions in the DNA examined to date (i.e., multiple genes/cosmid). Given this wide variation in gene density and current sequencing costs, we propose that newly targeted genomic regions should be analyzed first by a "Lewis and Clark" exploratory approach, before committing to full length DNA sequencing. We are using a SAmple SEquencing (SASE) approach to rapidly generate aligned sequences along the chromosome 5 and 16 physical maps. SASE analysis is a method for rapidly "scanning" large genomic regions with minimal cost, identifying, and localizing most genes. Briefly, individual cosmids are partially digested with Sau3A and 3 kb fragments are recloned into double-strand sequencing vectors. By sequencing both ends of a 1X sampling of these recloned fragments along with end sequences of the cosmid, 70% sequence coverage is achieved with 98% clone coverage. The majority of this clone coverage is ordered by the relationship between the subclone end sequences. These ordered sequences are ideal substrates for directed sequencing strategies (for example, primer walking or transposon sequencing). SASE analysis has been initiated on the 40 Mb short arm of chromosome 16 and the 45 Mb short arm of chromosome 5. We propose to make SASE sequences, along with feature annotation, publicly available through GSDB. Such data are sufficient to allow PCR amplification of the sequenced region from GSDB submissions alone, eliminating the need for extensive clone archiving and distributing, will allow for the effective "democratization" of the genome, allowing numerous laboratories to share and contribute to the growing genome databases.

DOE Grant No. DE-FG03-96ER62298.

One-Step PCR Sequencing

Kenneth W. Porter, J. David Briley, and **Barbara Ramsay Shaw**
Department of Chemistry; Duke University; Durham, NC 27708
919/660-1553, Fax: -1605, *ken@chem.duke.edu*

A method is described to simultaneously amplify and sequence DNA using a new class of nucleotides containing boron. During the polymerase chain reaction, boron-modified nucleotides, i.e. 2'-deoxynucleoside 5'-a-[P-borano]-triphosphates,[1,2] are incorporated into the product DNA. The boranophosphate linkages are resistant to nucleases and thus the positions of the boranophosphates can be revealed by exonuclease digestion, thereby generating a set of fragments that defines the DNA sequence. The boranophosphate method offers an alternative to current PCR sequencing methods.

Single-sided primer extension with dideoxynucleotide chain terminators is avoided with the consequence that the sequencing fragments are derived directly from the original PCR products. Boranophosphate sequencing is demonstrated with the Pharmacia and the Applied Biosystems 373A automatic sequencers producing data that is comparable to cycle sequencing.

DOE Grant No. DE-FG02-97ER62376 and NIH Grant No. HG00782.

References

[1] Sood, A., Shaw, B. R., and Spielvogel, B. F. (1990) J. Amer. Chem. Soc. 112, 9000-9001.
[2] Tomasz, J., Shaw, B. R., Porter, K., Spielvogel, B. F., and Sood, A. (1992) Angew. Chem. Int. Ed. Engl. 31. 1373-1375.

Automation of the Front End of DNA Sequencing

Lloyd M. Smith and **Richard A. Guilfoyle**
University of Wisconsin; Madison, WI 53706
Guilfoyle: 608/265-6138, Fax: -6780
raguilfo@facstaff.wisc.edu

The objective of this project is to continue developing more efficient tools and methods addressing the "front-end" processes of large-scale DNA sequencing. Our specific aims are high-throughput purification and mapping of cosmid inserts, controlled fragmentation of random inserts, direct selection vectors for cloning and sequencing, high-throughput M13 clone isolations, and high-throughput template purifications.

An approach to multi-cosmid purifications was developed using a cell-harvester and binding to GF/C glass fiber filter-bottom microtiter plates. This method proved inadequate because the yields were low and the DNA was easily fragmented. In the last year we have started examining the use of triplex-affinity capture (TAC) for this purpose as applied to BACs, based on our previous success with TAC purification and restriction mapping of cosmids (1,2).

We initially proposed to control random fragmentation for shotgun cloning using CviJ1 and its methyltransferase. Instead, we are now exploring automating it by scaled-down nebulization and parallel processing.

We have made a vector, M13-102 (3,4, patented)), for facilitating construction and improving quality of M13 shotgun libraries. It allows direct selection of recombinants, dephosphorylation of inserts to reducing chimerics, contains universal primers for fluorescent sequencing, and a triplex sequence for easy TAC purification of linearized RF DNA. We also made a version of this vector, M13-100Z, which expressed the alpha-peptide of B-gal. Its utility is in flow cytometry based clone isolation. We continue to develop these vectors for multiple cloning sites, and insert flipping using in closing steps of large-scale sequencing projects.

We continue to develop high-throughput clone isolations by flow cytometric cell sorting. M13 or plasmid clones can theoretically be isolated at rates in microtiter wells at rates up to 2 per second using our present FacStar-Plus cytometer and collection assembly. Theoretical rates are much higher. This bypasses plating onto solid-media and any need for plaque/colony picking. We initially tried isolations after microencapsulation of cells in agarose gel microbeads, but with H/W and S/W improvements we can now distinguish positively selected transfected cells from background. Efficiency of sorting is very sensitive to detection efficiency. We continue to investigate different methods of florescence detection for various plasmid and M13 vector systems including fluorogenic substrates for B-gal, fluorescent-tagged antibodies to M13 or cell surface proteins, and green fluorescent protein as a reporter.

We have been developing a solid-phase filter plate method for M13 template purifications using carboxylated polystyrene beads (Bangs Labs, IN) for automating on the Hamilton 2200. It should process 96 samples in under 30 minutes and deliver 1-2 micrograms per sample for cycle-sequencing. This approach has proven superior to others we have tried with respect to amenability to automation (5,6).

Ancillary projects. We reported a method for direct fluorescence analysis of genetic polymorphisms using oligonucleotide arrays on glass supports (7), which spun off other projects including (a) enhanced discrimination by artificial mismatch hybridization (8), restriction hybridization ordering of shotgun clones, and restriction site indexing-PCR (RSI-PCR) (9, patent applied for). RSI-PCR is an alternative strategy to extra-long PCR which has application in large gap filling (>45kb) differential gene expression analysis, RFLP and EST marker production, end-sequencing and others.

Our most significant findings are the following:

1. Improved direct selection M13 cloning vector
2. Rapid restriction mapping of cosmids using triple-helix affinity capture
3. High-throughput M13 template production using carboxylated beads
4. Sequencing of a cosmid encoding the Drosophila GABA receptor
5. Improved detection of sequencing clones by flow-cytometry
6. RSI-PCR, a strategy to obtain mapped and sequence-ready DNA directly from up to 0.5 kb regions of a complex genome using palindromic class II restriction enzymes; bypasses conventional cloning methodology (see previous section for applications).

DOE Grant No. DE-FG02-91ER61122.

References

1. Ji, H., Smith, L.M., and Guilfoyle, R.A. (1994) GATA 11, 43-47.
2. Ji, H., Francisco, T., Smith, L.M. and Guilfoyle, R.A. (1996) Genomics 31, 185-192.
3. Guilfoyle,R. and Smith, L.M. (1994) Nucleic Acids Res. 22, 100-107.
4. Chen, D., Johnson, A.F., Severin, J.M., Rank, D.R., Smith, L.M. and Guilfoyle, R.A. (1996) Gene 172, 53-57.
5. Kolner, D.E., Guilfoyle, R.A., and Smith, L. (1994) DNA Sequence 4, 253-257.
6. Johnson, A.F., Wang, R., Ji, H., Chen, D., Guilfoyle, R.A. and Smith, L.M. (1996) Anal Biochem 234, 83-95.
7. Guo, Z., Guilfoyle, R.A., Thiel, A.J., Wang, R. and Smith, L.M. (1994) Nucleic Acids Res, 22, 5456-5465.
8. Guo, Z., Liu, Q., and Smith, L.M. (submitted).
9. Guilfoyle, R.A., Guo, Z., Kroening, D., Leeck, C. and Smith, L.M.(submitted).

High-Speed DNA Sequence Analysis by Matrix-Assisted Laser Desorption Mass Spectrometry

Lloyd M. Smith and Brian Chait[1]
Department of Chemistry; University of Wisconsin; Madison, WI 53706
608/263-2594, Fax: /265-6780, *smith@chem.wisc.edu*
[1]Rockefeller University; New York, NY 10021

Our mass spec research has focused primarily on the possibility of utilizing Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI-MS) as an alternative method to conventional gel electrophoresis for DNA sequence analysis. In this approach, extension fragments generated by the Sanger sequencing reactions are separated by size and detected in the mass spectrometer in one step.

Our group has shown fragmentation to be a major factor limiting accessible mass range, sensitivity, and mass resolution in the analysis of DNA by MALDI-MS. This DNA fragmentation was shown to be strongly dependent on both the MALDI matrix and the nucleic acid sequence employed. Fragmentation is proposed to follow a pathway in which nucleobase protonation leads to cleavage of the N-glycosidic bond with base loss, followed by cleavage of the phosphodiester backbone. Modifications of the deoxyribose sugar ring by replacing the 2' hydrogen with more electron-withdrawing groups such as the hydroxyl or fluoro group were shown to stabilize the N-glycosidic bond, partially or completely blocking fragmentation at the modified nucleosides. The stabilization provided by these chemical modifications was also shown to expand the range of matrices useful for nucleic acid analysis, yielding in some cases greatly improved performance.

DOE Grant No. DE-FG02-91ER61130.

Relevant Publication

Zhu, L.; Parr, G. P.; Fitzgerald, M. C.; Nelson, C. M.; Smith, L. M. Oligodeoxynucleotide fragmentation in MALDI/TOF Mass spectrometry using 355 nm radiation. J. Am. Chem. Soc. 1995, 117, 6048-6056.

Analysis of Oligonucleotide Mixtures by Electrospray Ionization-Mass Spectrometry

Richard D. Smith, David C. Muddiman, James E. Bruce, and Harold R. Udseth
Environmental Molecular Sciences Laboratory; Pacific Northwest National Laboratory; Richland, WA 99352
509/376-0723, Fax: -5824, *rd_smith@pnl.gov*
http://www.emsl.pnl.gov:2080/docs/msd/fticr/advmasspec.html

This project aims to develop electrospray ionization mass spectrometry (ESI-MS) methods for high speed DNA sequencing of oligonucleotide mixtures, that can be integrated into an effective overall sequencing strategy. A second goal is develop mass spectrometric methods that can be effective utilized in post genomic research in broad areas of DNA characterization, such as with polymerase chain reaction to rapidly and accurately identify single base polymorphisms. ESI produces intact molecular ions from DNA fragments of different size and sequence with high efficiency [1]. Our aim is to determine ESI mass spectrometry conditions that are compatible with biological sample preparation to allow efficient ionization of DNA and allowing for the analysis of complex mixtures (e.g., Sanger sequencing ladder). We have developed a novel on-line microdialysis method at PNNL to remove salts, detergents, and buffers from such biological preparations as PCR and dideoxy sequencing mixtures. This has allowed for rapid and efficient desalting (e.g., of samples having 0.25 M NaCl) allowing ESI mass spectral analysis without the typically problematic Na-adducts observed. Oligonucleotide ions are typically produced from ESI with a broad distribution of net charge states for each molecular species, and thus leading to difficulties in analysis of complex mixtures [1]. To make identification of each component in a sequencing mixture possible, the charge states of molecular ions can be reduced using gas-phase reactions. The charge-state reduction methods being examined include: (1) reactions with organic acids and bases (in the solution to be electrosprayed and the ESI-MS interface or the gas phase); (2) the labeling of the oligonucleotides with a designed functional group for production of molecular ions of very low charge states; and (3) the shielding of potential charge sites on the oligonucleotide phosphate/phosphodiester groups with polyamines (and the subsequent gas-phase removal of the neutral amines). In initial studies two methods for charge state reduction of gas phase oligonucleotide negative ions have been tested: (1) the addition of acids and bases to the oligonucleotide solution and (2) the formation of diamine adducts followed by dissociation in the interface region [2,3]. Several methods show promise for charge state reduction and results have been demonstrated for series of smaller oligonucleotides. We have recently demonstrated for the first time that PCR products can be rapidly detected using ESI-MS with significant improvements projected [4,5]. Finally, new mass spectrometric methods have been developed to provide the dynamic range expansion necessary for addressing DNA sequencing mixtures [6]. Our overall aim is to provide a foundation for the development of an overall approach to high speed sequencing (including the rapid and precise PCR product characterization) using cost effective high-throughput instrumentation.

DOE Contract No. DE-AC06-76RLO-1830.

References

[1] "New Developments in Biochemical Mass Spectrometry: Electrospray Ionization", R. D. Smith, J. A. Loo, C. G. Edmonds, C. J. Barinaga, and H.R. Udseth, Anal. Chem., 62, 882-889 (1990).

[2] "Charge State Reduction of Oligonucleotide Negative Ions from Electrospray Ionization", X. Cheng, D. C. Gale, H. R. Udseth, and R. D. Smith, Anal. Chem., 67, 586-593 (1995).

[3] "Charge-State Reduction with Improved Signal Intensity of Oligonucleotides in Electrospray Ionization Mass Spectrometry" D.C. Muddiman, X.Cheng, H.R. Udseth and R.D. Smith J. Am. Soc. Mass Spectrom., 7 (8) 697-706 (1996).

[4] "Analysis of Double-stranded Polymerase Chain Reaction Products from the Bacillus cereus Group by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry" D.S. Wunschel, K.F. Fox, A. Fox, J.E. Bruce, D.C. Muddiman and R.D. Smith Rapid Commun. in Mass Spectrom., 10, 29-35 (1996).

[5] "Characterization of PCR Products From Bacilli Using Electrospray Ionization FTICR Mass Spectrometry", D. C. Muddiman, D. S. Wunschel, C. Liu, L. Pasa-Tolic, K. F. Fox, A. Fox, G. A. Anderson, and R. D. Smith, Anal. Chem., 68, 3705-3712 (1996).

[6] "Colored Noise Waveforms and Quadrupole Excitation for the Dynamic Range Expansion in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry", J. E. Bruce, G. A. Anderson and R. D. Smith, Anal. Chem., 68, 534-541 (1996).

High-Speed Sequencing of Single DNA Molecules in the Gas Phase by FTICR-MS

Richard D. Smith, David C. Muddiman, S. A. Hofstadler, and J. E. Bruce
Environmental Molecular Sciences Laboratory; Pacific Northwest National Laboratory; Richland, WA 99352
509/376-0723, Fax: -5824, *rd_smith@pnl.gov*
http://www.emsl.pnl.gov:2080/docs/msd/fticr/advmasspec.html

This project is aimed at the development of a totally new concept for high speed DNA sequencing based upon the analysis of single (i.e., individual)large DNA fragments using electrospray ionization (ESI) combined with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. In our approach, large single-stranded DNA segments extending to as much as 25 kilobases (and possibly much larger), are transferred to the gas phase using ESI. The multiply-charged molecular ions are trapped in the cell of an FTICR mass spectrometer, where one or more single ion(s) are then selected for analysis in which its mass-to-charge ratio (m/z) is measured both rapidly and non-destructively. Single ion detection is achievable due to the high charge state of the electrosprayed ions and the unique sensitivity of new FTICR detection methodologies.

Initial efforts under this project have demonstrated the capability for the formation, extended trapping, isolation, and monitoring of sequential reactions of highly charged DNA molecular ions with molecular weights well into the megadalton range [1-6]. We have shown that large multiply-charged individual ions of both single and double-stranded DNA anions can also be efficiently trapped in an FTICR cell, and their mass-to-charge ratios measured with very high accuracy. Thus, it is feasible to quickly determine the mass of each lost unit as the DNA is subjected to rapid reactive degradation steps. One approach is to develop methods based upon the use of ion-molecule or photochemical processes that can promote a stepwise reactive degradation of gas-phase DNA anions. Successful development of one of these approaches could greatly reduce the cost and enhance the speed of DNA sequencing, potentially allowing for sequencing DNA segments of more than 25 kilobase in length, on a time scale of minutes with negligible error rates with the added potential for conducting many such measurements in parallel. Instrumentation optimized for these purposes is currently being introduced and promises to greatly advance the methodology. The techniques being developed promise to lead to a host of new methods for DNA characterization, potentially extending to the size of much larger DNA restriction fragments (>500 kilobases).

DOE Contract No. DE-AC06-76RLO-1830.

(abstract continued)

References

[1] "Trapping Detection and Reaction of Very Large Single Molecular Ions by Mass Spectrometry," R. D. Smith, X. Cheng, J. E. Bruce, S.A. Hofstadler and G.A. Anderson, Nature, 369, 137-139 (1994).

[2] "Charge State Shifting of Individual Multiply-Charged Ions of Bovine Albumin Dimer and Molecular Weight Determination Using an Individual-Ion Approach," X. Cheng, R. Bakhtiar, S. Van Orden, and R. D. Smith, Anal.Chem., 66, 2084-2087 (1994).

[3] "Trapping, Detection, and Mass Measurement of Individual Ions in a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer,: J.E. Bruce,X. Cheng, R. Bakhtiar, Q. Wu, S.A. Hofstadler, G.A. Anderson, and R.D.Smith, J. Amer. Chem. Soc., 116, 7839-7847 (1994).

[4] "Direct Charge Number and Molecular Weight Determination of Large Individual Ions by Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance Mass Spectrometry", R. Chen, Q. Wu, D.W. Mitchell, S.A. Hofstadler, A.L. Rockwood, and R. D. Smith, Anal. Chem., 66, 3964-3969 (1994).

[5] "Trapping, Detection and Mass Determination of Coliphage T4 (108 MDa) Ions by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry" R. Chen, X. Cheng, D.W. Mitchell, S.A. Hofstadler, A.L. Rockwood, Q. Wu, M.G. Sherman and R.D. Smith, Anal. Chem.,67, 1159-1163 (1995).

[6] "Accurate Molecular Weight Determination of Plasmid DNA Using Mass Spectrometry", X. Cheng, D. G. Camp II, Q. Wu, R. Bakhtiar, D. L. Springer, B.J. Morris, J. E. Bruce, G. A. Anderson, C. G. Edmonds and R. D. Smith, Nucleic Acid Res., 24, 2183-2189 (1996).

Characterization and Modification of DNA Polymerases for Use in DNA Sequencing

Stanley Tabor
Harvard University; Boston, MA 02115-5730
617/432-3128, Fax: -3362, *tabor@bcmp.med.harvard.edu*
http://sbweb.med.harvard.edu/~bcmp
http://sbweb.med.harvard.edu/~bcmp/tabor.html

Our studies are directed towards improving the properties of DNA polymerases for use in DNA sequencing. The primary focus is understanding the mechanism by which DNA polymerases discriminate against nucleotide analogs, and the mechanism by which they incorporate nucleotides processively without dissociating from the DNA template.

We are comparing three DNA polymerases that have been used extensively for DNA sequencing; *E. coli* DNA polymerase I, T7 DNA polymerase, and Taq DNA polymerase. These are related to one another, and this homology has been exploited to construct active site hybrids that have been used to determine the structural basis for differences in their activities. Specifically, the hybrids have been used (1) to determine why *E. coli* DNA polymerase I and Taq DNA polymerase discriminate strongly against dideoxynucleotides, and (2) to understand how T7 DNA polymerase interacts with its processivity factor, thioredoxin, to confer high processivity.

Based on these studies, we have been able to modify Taq DNA polymerase and *E. coli* DNA polymerase I to make them incorporate dideoxynucleotides much more efficiently, and to have increased processivity in the presence of thioredoxin. The ability to incorporate dideoxynucleotides efficiently greatly improves the uniformity of band intensities on a DNA sequencing gel, thereby increasing the accuracy of the DNA sequence obtained. In addition, the efficient use of dideoxynucleotides reduces the amount of these analogs required for DNA sequencing, an important issue when using fluorescently modified dideoxy terminators. In an approach that complements these studies, we, in collaboration with Dr. Thomas Ellenberger (Harvard Medical School), are determining the crystal structure of T7 DNA polymerase in a complex with thioredoxin and a primer-template. Knowledge of this structure will allow the rationale design of specific mutations that will enable DNA polymerases to incorporate other analogs useful for DNA sequencing more efficiently, such as those with fluorescent moieties on the bases.

DOE Grant No. DE-FG02-96ER62251.

Relevant Publication

Tabor, S., and Richardson, C. C. (1995). A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy-and dideoxyribonucleotides. Proc. Natl. Acad. Sci. U.S.A. 92, 6339-6343.

Bedford, E., Tabor, S. and Richardson, C. C. (1997). The thioredoxin binding domain of bacteriophage T7 DNA polymerase confers processivity on *Escherichia coli* DNA polymerase I. Proc. Natl. Acad. Sci. U.S.A. 94, 479-484.

Modular Primers for DNA Sequencing

Mugasimangalam Raja,[1,2] Dina Sonkin,[2] Lev Lvovsky,[2] and **Levy Ulanovsky**[1,2]
[1]Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory, Argonne, IL 60439-4833
Ulanovsky: 630/252-3940; Fax: -3387, *levy@anl.gov*
[2]Dept. of Structural Biology; Weizmann Institute of Science; Rehovot 76100, Israel

We are developing molecular approaches to DNA sequencing enabling primer walking without the step of chemical synthesis of oligonucleotide primers between the walks. One such approach involves "modular primers" described earlier, consisting of 5-mers, 6-mers or 7-mers (selected from a presynthesized library), annealing to the template contiguously with each other. Another approach, that we have termed DENS (Differential Extension with Nucleotide Subsets), works by selectively extending a short primer, making it a long one at the intended site only. DENS starts with a limited initial extension of the primer (at 20-30 C) in the presence of only 2 out of the 4 possible dNTPs. The primer is extended by 6-9 bases or longer at the intended priming site, which is deliberately selected, (as is the two-dNTP set), to maximize the extension length. The subsequent sequencing/termination reaction at 60-65 C then accepts the extended primer at the intended site, but not at alternative sites, where the initial extension (if any) is generally much shorter. DENS allows the use of primers as long as 8-mers (degenerate in 2 positions) which prime much more strongly than modular primers involving 5-7 mers and which (unlike the latter) can be used with thermostable polymerases, thus allowing cycle-sequencing with dye-terminators for Taq, as well as making double-stranded DNA sequencing more robust.

These technologies are expected to speed up genome sequencing in more than one way:

a) Reduction in redundancy would result from more efficient and rapid closure of even long gaps which are currently avoided at the price of 7-to 9-fold redundancy in shotgun. Instantly available primers would also improve the quality of sequencing. Stretches of sequence that have too low confidence level (high suspected error rate) can be resequenced without synthesizing new oligos and without growing any new subclones.

b) Further down the road, the completion of the automation of the closed cycle of primer walking will be made possible via the elimination of the need to synthesize the walking primers. Combined with the capillary sequencers, the instant availability of the walking primers should reduce the time per walking cycle from 2-3 days now to about 1.5-2.0 hours, an improvement in speed by a factor of 20-50.

c) The closed-end automation would minimize both the labor cost and human errors. As primer walking has minimal, if any, front-end and back-end bottlenecks inherent to shotgun, the cost of sequencing would be essentially that of reagents, 5 cents/base or less.

DOE Grant No. DE-FG02-94ER61831.

Time-of-Flight Mass Spectroscopy of DNA for Rapid Sequence

Peter Williams, Chau-Wen Chou, David Dogruel, Jennifer Krone, Kathy Lewis, and Randall Nelson
Department of Chemistry and Biochemistry; Arizona State University; Tempe, AZ 85287
602/965-4107, Fax: -2747, *pw@asu.edu*

There are three potential roles for mass spectrometry relevant to the Human Genome Project:

a) The most obvious role is that on which all groups have been focussing -development of an alternative, faster sequence ladder readout method to speed up large-scale sequencing. Progress here has been difficult and slow because the mass spectrometry requirements exceed the current capabilities of mass spectrometry even for proteins, and DNA presents significantly more difficulty than proteins. We have shown previously that pulsed laser ablation of DNA from frozen aqueous films has the potential to yield sequence-quality mass spectra, but that ionization in this approach is erratic and uncontrollable. We are focussing on developing ionization methods using ion (or electron) attachment to vapor-phase DNA (ablated from ice films) in an electric field-free environment; results of this approach will be reported.

b) Mass spectrometry may not ultimately compete favorably in speed with large-scale multiplexing of conventional or near-term technologies such as capillary electrophoresis. However, as the Genome project nears completion there will be an increasing need for rapid small-scale DNA analysis, where the multiplex advantage will not be so great and mass spectrometry could play a more significant role there. With this in mind we are looking at ways to speed up the overall mass spectrometric analysis, e.g. simple rapid cleanup of sequence mixtures, and at generation of short sequence ladders by exopeptidase digestion.

c) Given the genome data base(s) at the completion of the project, with rapid search capability, a need will arise for comparably rapid generation of search input data to identify often very small quantities of proteins isolated from biochemical investigations. With this in mind we have developed extremely rapid enzyme digestion techniques optimized for mass spectrometric readout, using endopeptidases covalently coupled directly to the mass spectrometer probe tip. The elimination of autolysis and transfer losses allows rapid (few minute) endopeptidase digestion and mass analysis of as little as 1 picomole of protein, leading to an ambiguous database identification. An alternative search procedure uses partial amino-acid sequence information. With the added use of exopeptidases to generate a peptide ladder sequence in the mass spectrum of the endopeptidase digest, on the order of a dozen residues of internal sequence can be generated in a total analysis time of 20 minutes or less, again using only picomoles of sample.

DOE Grant No. DE-FG02-91ER61127.

Development of Instrumentation for DNA Sequencing at a Rate of 40 Million Bases Per Day

Edward S. Yeung, Huan-Tsung Chang, Qingbo Li, Xiandan Lu, and Eliza Fung
Ames Laboratory and Department of Chemistry; Iowa State University; Ames, IA 50011
515/294-8062, Fax: -0266, *yeung@ameslab.gov*

We have developed novel separation, detection, and imaging techniques for real-time monitoring in capillary electrophoresis. These techniques will be used to substantially increase the speed, throughput, reliability, and sensitivity in DNA sequencing applications in highly multiplexed capillary arrays. We estimate that it should be possible to eventually achieve a raw sequencing rate of 40 million bases per day in one instrument based on the standard Sanger protocol. We have reached a stage where an actual sequencing instrument with 100 capillaries can be built to replace the Applied Biosystems 373 or 377 instruments, with a net gain in speed and throughput of 100-fold and 24-fold, respectively.

The substantial increase in sequencing rate is a result of several technical advances in our laboratory. (1) The use of commercial linear polymers for sieving allows replaceable yet reproducible matrices to be prepared that have lower viscosity (thus faster migration rates) compared to polyacrylamide. (2) The use of a charge-injection device camera allows random data acquisition to decrease data storage and data transfer time. (3) The use of distinct excitation wavelengths and cut-off emission filters allows maximum light throughput for efficient excitation and sensitive detection employing the standard 4-dye coding. (4) The use of indexmatching and 1:1 imaging reduces stray light without sacrificing the convenience of on-column detection.

Continuing efforts include further optimization of the separation matrix, development of new column conditioning protocols, refinement of the excitation/emission optics, design of a pressure injection system for 96-well titer plates, validation of a new 2-color base-calling scheme, simplification of software to allow essentially real-time data processing, implementation of voltage programming to shorten the total run times, and scale up of the technology to allow parallel sequencing in up to 1,000 capillaries.

Relevant Publications

K. Ueno and E. S. Yeung, "Simultaneous Monitoring of DNA Fragments Separated by Capillary Electrophoresis in a Multiplexed Array of 100 Channels", Anal. Chem. 66, 1424-1431 (1994).

X. Lu and E. S. Yeung, "Optimization of Excitation and Detection Geometry for Multiplexed Capillary Array Electrophoresis of DNA Fragments", Appl. Spectrosc. 49, 605-609 (1995).

Q. Li and E. S. Yeung, "Evaluation of the Potential of a Charge Injection Device for DNA Sequencing by Multiplexed Capillary Electrophoresis", Appl. Spectrosc. 49, 825-833 (1995).

E. N. Fung and E. S. Yeung, "High-Speed DNA Sequencing by Using Mixed Poly(ethyleneoxide) Solutions in Uncoated Capillary Columns," Anal. Chem. 67, 1913-1919 (1995).

Q. Li and E. S. Yeung, "Simple Two-Color Base-Calling Schemes for DNA Sequencing Based on Standard 4-Label Sanger Chemistry", Appl. Spectrosc. 49, 1528-1533 (1995).

Resolving Proteins Bound to Individual DNA Molecules

David Allison, Bruce Warmack, Mitch Doktycz, Tom Thundat, and Peter Hoyt
Molecular Imaging Group; Health Sciences Research Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-6123
Allison: 423/574-6199, Fax: -6210, *allisondp@ornl.gov*
Warmack: 423/574-6202, Fax: -6210, *rjw@ornl.gov*

We have precisely located sequence specific proteins bound to individual DNA molecules by direct AFM imaging. Using a mutant *Eco*R I endonuclease that site-specifically binds but doesn't cleave DNA, bound enzyme has been imaged and located, with an accuracy of ±1%, on well characterized plasmids and bacteriophage lambda DNA (48 kb). Cosmids have been mapped and, by incorporating methods for anchoring molecules to surfaces and straightening to prevent molecular entanglement, BAC-sized clones could be analyzed.

This direct imaging approach could be rapidly developed to locate other sequence-specific proteins on genomic clones. Enzymatic proteins, involved in identifying and repairing damaged or mutated regions on DNA molecules, could be imaged bound to lesion sites. Transcription factor proteins that identify gene-start regions and other regulatory proteins that modulate the expression of genes by binding to specific control sequences on DNA molecules could be precisely located on intact cloned DNAs.

Conventional gel-based techniques for identifying site-specific protein binding sites must rely upon fragment analysis for identifying restriction enzyme sites, or, for non-cutting proteins, upon gel-shift methods that can only address small DNA fragments. Conversely, AFM imaging is a general approach that is applicable to the analysis of all site-specific DNA protein interactions on large-insert clones. This technique could be developed for high-throughput analysis, can be accomplished by technicians, uses readily available relatively inexpensive instrumentation, and should be a technology fully transferable to most laboratories.

DOE Contract No. DE-AC05-840R21400.

*Improved Cell Electrotransformation by Macromolecules

Alexandre S. Boitsov, Boris V. Oskin, Anton O. Reshetin, and Stepan A. Boitsov
Department of Biophysics; St.Petersburg State Technical University; 195251 St. Petersburg, Russia
+7-812/277-5959, Fax: /247-2088 or /534-3314,
sasha@bioph.hop.stu.neva.ru

Our work for 1996 and 1997 will include the following:

1. Comparative study of the kinetics of entry of DNA of different molecular forms into E.coli cells DH10B/r and DH5a during electrotransformation. Study of the optimal regimes of cell-wall permeabilization for the DH10B/r cells.

2. Study of the efficiency of BAC cloning in DH10B/r cells using new electrotransformation method. Optimization of the procedure for DH10B/r cells.

3. Modernization of the electronic equipment in accordance with results of the biological experiments. To expand the studies, we need to extend the capability of the instrumentation to increase its flexibility and to improve the accuracy and reproducibility of the electric fields we generate by incorporating electronic components with higher tolerances.

DOE Grant No. OR00033-93CIS015.

Overcoming Genome Mapping Bottlenecks

Charles R. Cantor
Center for Advanced Biotechnology; Boston University; Boston MA 02215
617/353-8500, Fax: 8501, *crc@enga.bu.edu*
http://eng.bu.edu/CAB

Most traditional DNA analysis is done based on fractionation of DNA by length. We have, instead, begun to explore the use of DNA sequences as capture and detection methods to expedite a number of procedures in genome analysis.

Triplet repeats like $(GGC)_n$ are an important class of human genetic markers, and they are also responsible for a number of inherited diseases involving the central nervous system. For both of these reasons it would be very useful to have a way to monitor the status of large numbers of triplet repeats simultaneously. We are developing methods to isolate and profile classes of such repeats.

In one method, genomic DNA is cut with one or more restriction nucleases, and splints are ligated onto the ends of the fragments. Then fragments containing a specific class of repeats are isolated by capture on magnetic microbeads containing an immobilized simple repeating sequence. The desired material is then released, and, if necessary, a selective PCR is done to reduce the complexity of the sample. Otherwise the entire captured sample is amplified by PCR. The spectrum of repeats is then examined by electrophoresis on an automated fluorescent gel reader. In our case the Pharmacia ALF is used, because of its excellent quantitative signal accuracy. A very complex spectrum of bands is seen representing hundreds of DNA fragments. We have shown that this spectrum is dramatically different with DNAs from unrelated individuals, and the spectrum is markedly dependent on the choice of restriction enzyme, as expected. Repeated measurements on the same sample are highly reproducible. The ability of the method to detect a specific altered repeat length in a complex DNA sample has been validated by examining several individuals with normal or expanded repeat sequences in the Huntington's disease gene. One very powerful application of this method may be the analysis of potential DNA differences in monozygotic twins discordant for a genetic disease. This method can be used to capture genome subsets containing any interspersed repeat. It will also detect insertions and deletions nearby such repeats. Methylation differences between sensitive methylation samples are also detectable when restriction fragments are used.

---

*Projects designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).

Conventional analysis of triplet repeats is very laborious since individual repeats must be analyzed by electrophoresis on DNA sequencing gels. The decrease in effort for such analyses will scale linearly as the number of repeats that can be analyzed simultaneously, so we are potentially looking at something like a factor of 100 improvement if the above scheme under development can be effectively realized.

As an alternative approach, we are developing chip-based methods that can detect the length of a tandemly-repeating sequence without any need for gel electrophoresis. Here the goal is to build an array of all possible repeat sequence lengths flanked by single-copy DNA. When an actual sample is hybridized to such an array, the specific alleles in the sample will produce perfect duplexes at their corresponding points in the array and at mismatched duplexes elsewhere. Thus, the task of scoring the repeat lengths is reduced to the task of distinguishing perfect and imperfect duplexes. Currently we are exploring a number of different enzymatic protocols that offer the promise of making such distinctions reliably.

In other work we are using enzyme-enhanced sequencing by hybridization (SBH) as a device for the rapid preparation of DNA samples for mass spectrometry. For example, partially duplex DNA probes can capture and generate sequence ladders from any arbitrary DNA sequence. Current MALDI protocols allow sequence to be read to lengths of 50 to 60 bases. While this is probably insufficient for most de novo DNA sequencing, it is an extremely promising approach for comparative or diagnostic DNA sequencing.

DOE Grant No. DE-FG02-93ER61609.

Preparation of PAC Libraries

Joe Catanese, Baohui Zhao, Eirik Frengen, Chenyan Wu, Xiaoping Guan, Chira Chen, Eugenia Pietrzak, Panayotis A. Ioannou,[1] Julie Korenberg,[2] Joel Jessee,[3] and **Pieter J. de Jong**
Department of Human Genetics; Roswell Park Cancer Institute; Buffalo, NY 14263
de Jong: 716/845-3168, Fax: -8849
pieter@dejong.med.buffalo.edu
http://bacpac.med.buffalo.edu
[1]The Cyprus Institute of Neurology and Genetics; Nicosia, Cyprus
[2]Cedars Sinai Medical Center; Los Angeles, CA 90048
[3]Life Technologies, Gaithersburg, MD 20898

Recently, we have developed procedures for the cloning of large DNA fragments using a bacteriophage P1 derived vector, pCYPAC1 (Ioannou et al. (1994), Nature Genetics 6: 84-89). A slightly modified vector (pCYPAC2) has now been used to create a 15-fold redundant PAC library of the human genome, arrayed in more than 1,000 384-well dishes. DNA was obtained from blood lymphocytes from a male donor. The library was prepared in four distinct sections designated as RPCI-1, RPCI-3, RPCI-4 and RPCI-5, respectively, each having 120 kbp average inserts. The RPCI-1 segment of the library (3X; 120,000 clones, including 25% non-recombinant) has been distributed to over 40 genome centers worldwide and has been used in many physical mapping studies, positional cloning efforts and in various large-scale DNA sequencing enterprises. Screening of the RPCI-1 library by numerous markers results in an average of 3 positive PACs per autosome-derived probe or STS marker. In situ hybridization results with 250 PAC clones indicate that chimerism is low or non-existing. Distribution of RPCI-3 (3X, 78,000 clones, less than 1% non-recombinants, 4% empty wells) is now underway and the further RPCI-4 and -5 segments (< 5% empty wells) will be distributed upon request. To facilitate screening of the PAC library, we have provided the RPCI-1 PAC library to several screening companies and noncommercial resource centers. In addition, we are now distributing high-density colony membranes at cost-recovery price, mainly to groups having a copy of the PAC library. The combined RPCI-1 and -3 segments (6X) can be represented on 11 colony filters of 22x22 cm, using duplicate colonies for each clone. We are currently generating a similar PAC library from the 129 mouse strain.

To facilitate the additional use of large-insert bacterial clones for functional studies, we have prepared new PAC & BAC vectors with a dominant selectable marker gene (the blasticidin gene under control of the beta-actin promoter), an EBV replicon and an "update feature". This feature utilizes the specificity of Transposon Tn7 for the Tn7att sequence (in the new PAC and BAC vectors) to transpose marker genes, other replicons and other sequences into PACs or BACs. Hence, it facilitates retrofitting existing PAC/BAC clones (made with the new vectors) with desirable sequences without affecting the inserts. The new vector(s) are being applied to generate second generation libraries for human (female donor), mouse and rat.

DOE Grant No. DE-FG02-94ER61883 and NIH Grant No. 1R01RG01165.

Development of Affinity Technology for Isolating Individual Human Chromosomes by Third-Strand Binding

Jacques R. Fresco and Marion D. Johnson III
Department of Molecular Biology; Princeton University; Princeton, NJ 08544-1011
609/258-3927, Fax: -6730
esteckman@molbiol.princeton.edu
http://molbiol.princeton.edu

Prior to the onset of this grant, solution conditions had been developed for binding a 17-residue third strand oligodeoxyribonucleotide probe to a specific human chromosome (HC) 17 multicopy alpha satellite target sequence cloned into DNA vectors of varying size up to 50 kb. Binding was shown to be both highly efficient and specific. Moreover, initial experiments with fluorescent-labeled third strands and human lymphocyte metaphase spreads and interphase nuclei proved similarly successful. During the current research period, the technology for such third strand-based cytogenetic examination, i.e., Triplex *In Situ* Hybridization or TISH, of such spreads was perfected, so that it is now a highly reproducible method. Comparison of spreads of different individuals by TISH and FISH analysis has provided a new basis for detecting alpha satellite DNA polymorphisms, the basis of which requires further investigation.

This year work also commenced on the development of comparable probes specific for alpha satellite sequences in HC-X, 11, and 16. The work with HC-X has reached the stage where we are ready to test the probe for TISH-based cytogenetic analysis. Solution studies of the interaction of the probes designed for HC-11 and HC-16 alpha satellite targets are following the well-established path we employed for HC-17 and HC-X. With the expectation of success in these cases during the coming year, the way should be clear for the development and application of comparable probes for alpha satellite sequences of any other human chromosomes that may be of interest, and possibly of other eukaryotic species.

Meanwhile, we have begun to turn our attention to two other goals, one being the exploitation of our probes for the isolation of individual human chromosomes by affinity purification, as we originally proposed. The other goal is to exploit our probes as aids in flow sorting human chromosomes, a direction of work we expect to pursue in collaboration with the Los Alamos National Laboratory, just as soon as they indicate a readiness to do so. Finally, we have begun to evaluate the possibility of using third-strand binding fluorescent probes for detection of single copy genes by means of photon counting, a goal which we plan to undertake with our colleague Robert Austin of our Physics Department.

DOE Grant No. DE-FG02-96ER622202.

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao
Eleanor Roosevelt Institute for Cancer Research; Denver, CO 80206
303/333-4515, Fax: -8423, *kao@eri.uchsc.edu*

The objective of this project is to construct and characterize chromosome region-specific libraries as resources for genome analysis. We have used our chromosome microdissection and MboI linker-adaptor technique (PNAS 88, 1844, 1991) to construct region-specific libraries for human chromosome 2 and other chromosomes. The libraries have been critically evaluated for high quality, including insert size, proportion of unique vs repetitive sequence microclones, percentage of microclones derived from dissected region, etc.

We have constructed and characterized 11 region-specific libraries for the entire human chromosome 2 (the second largest human chromosome with 243 Mb of DNA), including 4 libraries for the short arm and 6 libraries for the long arm, plus a library for the centromere region. The libraries are large, containing hundreds of thousands of microclones in plasmid vector pUC19, with a mean insert size of 200 bp. About 40-60% of the microclones contain unique sequences, and between 70-90% of the microclones were derived from the dissected region. In addition, we have isolated and characterized many unique sequence microclones from each library that can be readily sequenced as STSs, or used in isolating other clones with large inserts (like YAC, BAC, PAC, P1 or cosmid) for contig assembly. These libraries have been used successfully for high resolution physical mapping and for positional cloning of disease-related genes assigned to these regions, e.g. the cloning of the gene for hereditary nonpolypsis colorectal cancer (Cell 75, 1215, 1993).

For each library, we have established a plasmid sub-library containing at least 20,000 independent microclones. These sub-libraries have been deposited to ATCC for permanent maintenance and general distribution. The ATCC Repository numbers for these libraries are: #87188 for 2P1 library (region 2p23-p25, comprising 25 Mb); #87189 for 2P2 library (2p21-p23, 28 Mb); #87103 for 2P3 library (2p14-p16, 22 Mb); #87104 for 2P4 library (2p11-p13, 28 Mb); #77419 for 2Q1 library (2q35-q37, 28 Mb); #87308 for 2Q2 library (2q33-q35, 24 Mb); #87309 for 2Q3 library (2q31-q32, 26 Mb); #87310 for 2Q4 library (2q23-q24, 19 Mb); #87409 for 2Q5 library (2q21-q22, 23 Mb); #87410 for 2Q6 lbrary (2q11-q14, 31 Mb); and #87411 for 2CEN library (2p11.1-q11.1, 4 Mb). Details of these libraries have been described: Hum. Genet. 93, 557, 1994 (for 2P1 library); Cytogenet. Cell Genet. 68, 17, 1995 (for 2P2 library); Somat. Cell Mol. Genet. 20, 353, 1994 (for 2P3 library); Somat. Cell Mol. Genet. 20, 133, 1994 (for 2P4 library); Genomics 14, 769, 1992 (for 2Q1 library; Somat. Cell Mol. Genet. 21, 335, 1995 (for 2Q2, 2Q3 & 2Q4 libraries); Somat. Cell Mol. Genet. 22, 57, 1996 (for 2Q5, 2Q6 & 2CEN libraries).

Region-specific libraries and short insert microclones for chromosome 2 are particularly useful resources for its eventual sequencing because this chromosome is less exploited and detailed mapping information is lacking. We have also constructed 3 region-specific libraries for the entire chromosome 18 using similar methodologies, including 18P library (18p11.32-p11.1, 22 Mb); 18Q1 library (18q11.1-q12.3, 25 Mb); and 18Q2 library (18q21.1-q23, 34 Mb). Details of these libraries have been described (Somat. Cell Mol. Genet. 22, 191-199, 1996).

DOE Grant No. DE-FG03-94ER61819.

*Identification and Mapping of DNA-Binding Proteins Along Genomic DNA by DNA-Protein Crosslinking

V.L. Karpov, O.V. Preobrazhenskaya, S.V. Belikov, and D.E. Kamashev
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 17984, Russia
Fax: +7-095/135-1405, *karpov@genom-II.eimb.rssi.ru*

In 1995-1996 we continued to map and identify nonhistone proteins binding at loci along the yeast chromosome. Using DNA-protein crosslinking in vivo, we detected two polypeptides that probably correspond to core subunits of yeast RNA-polymerase II in the coding region of the transketolase gene (TKL2). Several nonhistone proteins were detected that bind to the upstream region of TKL2 and to an intergenic spacer between calmodulin (CMD1) and mannosyl transferase (ALG1) genes. The apparent molecular weight of these proteins was estimated. We also developed a new method to synthesize strand-specific probes.

Using DNA-protein crosslinking in vitro, we found the amino acid residues of the Lac-repressor that interacts with DNA. Only Lys-33 crosslinks with the Lac-operator in the specific complex. In addition to Lys-33, the N-terminal end of the protein also crosslinks in a nonspecific complex. Our results demonstrate that, in the presence of an inducer, the repressor's N-termini crosslink to the operator's outermost nucleotides. We suggest that binding of an inducer changes the orientation of the DNA-binding domain of the Lac repressor to the opposite of that found for the specific complex.

We plan to use a new method to increase resolution and thus identify amino acids and nucleotides that participate in DNA-protein recognition. The mechanisms of transcription regulation of some yeast genes will thus be further elucidated. Our approaches are based on DNA-protein crosslinking. Detailed analysis will be done for specific and nonspecific complexes, in the presence and absence of inducers. This will allow us to make some conclusions about possible conformational rearrangements in DNA-protein complexes during gene activation at the protein's DNA-binding domains.

DOE Grant No. OR00033-93C1S007.

References

1. Papatsenko D.A., Belikov S.V., Preobrazhenskaya O.V., and Karpov V.L. Two-dimensional gels and hybrydization for studying DNA-protein contacts by crosslinking // Methods in Molecular and Cellular Biology. 1995. V. 5, No 3. P.171-177.
2. Kamashev D., Esipova N.G., Ebralidse K., and Mirzabekov, A.D. Mechanism of lac repressor switch-off: Orientation of lac repressor DNA-binding domain is reversed upon inducer binding //FEBS Lett. 1995. V.375. P.27-30
3. Papatsenko D.A., Priporova I.V., Belikov S.V., and Karpov, V.L. Mapping of DNA-binding proteins along yeast genome by UV-induced DNA-protein crosslinking.// FEBS Letters, 1996, 381, 103-105.
4. Belikov S.V., Papatsenko D.A., and Karpov V.L. A method to synthesize strand-specific probes. //Anal.Biochemistry, 1996, 240,152-154.

A PAC/BAC Data Resource for Sequencing Complex Regions of the Human Genome: A 2-Year Pilot Study

Julie R. Korenberg
Cedars Sinai Medical Center; University of California;
Los Angeles, CA 90048-1869
310/855-7627, Fax: /652-8010
jkorenberg@mailgate.csmc.edu

While the complete sequencing the human genome at 99.99% accuracy is an immediate goal of the Human Genome Project, a serious technical deficiency remains the ability to rapidly and efficiently construct sequence ready maps as sequencing templates. This is particularly problematic in regions with unusual genome structure. An understanding of these troublesome regions prior to genome-wide sequencing will provide quality assurance as well as reliable sequencing strategies in these regions.

This proposal will generate a "whole genome" data resource to enable rapid and reliable sequencing of genomic DNA by the definition and characterization of the more than 52 regions of high homology now known to be distributed within unrelated genomic regions and cloned in BACs and PACs. To do this, we will:

1. Define regions of true homology in the human genome by characterizing subsets of the 4,700 BAC/PACs that generate multiple hybridization signals using fluorescence in situ hybridization (FISH). Of the 1,200 sites of multiple signals, more than 52 regions contain repeats as defined by 600 BAC/PACs. The chimerism rate, multiple clone wells, and chromosome of origin will be defined by re-streaking each clone, followed by fingerprint, FISH and PCR-based end-sequence analyses on hybrid panels and radiation hybrids.

Data will be shared with large sequencing efforts, deposited in the 4D database, available with annotation on ftp server and through GDB.

2. Generate contigs of BACs and PACs in regions of complex genome organization. Using STS, EST analyses, fingerprinting, BAC/PAC to BAC/PAC Southerns, end sequence walking in 3.5-20X libraries, and metaphase/interphase FISH, contigs will be seeded in 2-5 of the regions of known genome complexity, each of which is estimated as 2-5 Mb. These data will be used to evaluate and provide independent quality assurance of the STS and Radiation hybrid, and genetic maps in these regions. The most significant of these include 1p36/1q; 2p/q; multiple sites; 8p23 and 8 further sites; 9p/q.

3. Define additional regions of complex genomic structure. Library screening using known members of multiple member retro-transposon and other known repeated sequences defined by the ncbi database, followed by FISH analyzes to determine structure and potential large regions of associated homologies.

Collaboration with other genome and sequencing centers will provide quality control in the generation of sequence-ready maps for sequencing templates.

We believe that this effort is important since 1) it will provide a critical mapping tool necessary for the generation of sequence ready maps; 2) if initiated now, the problem areas could be delineated before scale ups to full production occur in major genome centers; 3) represents a modest cost such that the cost of these data would comprise only a small fraction of the cost of the entire genome sequence and would vastly decrease the cost of sequencing errors 4) and could be completed in a, short time (2 to 3 years) so as to be of maximum benefit to sequencing centers. The Principal Investigator in this project is ideally suited for this effort because the group has developed the technology and initiated FISH and genome analyses of over 4000 clones. We believe that this project represents a critical and timely effort to enable rapid and cost effective human genome sequencing.

Subcontract under Glen Evans' DOE Grant No. DE-FC03-96ER62294.

Mapping and Sequencing of the Human X Chromosome

D. L. Nelson, E.E. Eichler, B.A. Firulli, Y. Gu, J. Wu, E. Brundage, A.C. Chinault, M. Graves, A. Arenson, R. Smith, E.J. Roth, H.Y. Zoghbi, Y. Shen, M.A. Wentland, D.M. Muzny. J. Lu, K Timms, M. Metzger, and R.A. Gibbs
Department of Molecular and Human Genetics and Human Genome Center; Baylor College of Medicine; Houston, TX 77030
713/798-4787, Fax: -6370 or -5386, *nelson@bcm.tmc.edu*
http://www.bcm.tmc.edu/molgen

The human X chromosome is significant from both medical and evolutionary perspectives. It is the location of several hundred genes involved in human genetic disease, and has maintained synteny among mammals; both of these aspects are due to its role in sex determination and the haploid nature of the chromosome in males. We have addressed the mapping of this chromosome through a number of efforts, ranging from long-range YAC-based mapping to genomic sequence determination.

YAC mapping. The YAC-based map of the X is essentially complete. We have constructed a 40 Mb physical map of the Xp22.3-Xp21.3 region, spanning an interval from the pseudoautosomal boundary (PABX) to the Duchenne muscular dystrophy gene. This region is highly annotated, with 85 breakpoints defining 53 deletion intervals, 175 STSs (20 of which are highly polymorphic), and 19 genes.

Cosmid binning. The YAC-based physical is being used in a systematic effort to identify and sort cosmids prepared at LLNL from flow sorted X chromosomes into intervals. Gene identification through use of a common database for cDNA pool hybridization data is continuing. Over 50 YACs have been utilized as probes to the gridded cosmic arrays. These have identified over 9000 cosmids from the 24,000 member library. An additional 4000 cosmids have been identified using a variety of probes, with the bulk coming from cDNA pool probes. More recent emphasis has been placed on BAC clones as their identity for sequencing has been established. These have been identified using the usual methods.

Cosmid contig construction. Creation of long-range continuity in cosmids and BACs proceeds from clones identified by the YAC-based binning experiments. Identification of STS carrying clones is carried out by a combined PCR/

hybridization protocol, and adds to the specificity of the overlap data. Cosmids are grown and DNA is prepared by an Autogen robot. DNAs are digested and analyzed by the AB362 GeneScanner for collection of fingerprint data. The use of novel fluorescent dyes (BODIPY) in this application has increased signal strength markedly. End fragment detection is currently carried out with traditional Southern hybridization, however additional dyes will permit detection without hybridization in the GeneScanner protocol. Data are transferred to a Sybase database and analyzed with ODS (J. Arnold, U. Georgia) software for overlap. ODS output is ported to GRAM (LANL) for map construction. A fully automated approach has yet to be achieved, but this goal is increasingly in reach.

Sequencing. An independently funded project awarded to RAG seeks to develop long-range genomic sequence for ~2 Mb of the human X chromosome. In support of this project, cosmids have been constructed and isolated for the 1.6 Mb region between FRAXA and FRAXF in Xq27.3-Xq28. To date, the complete sequences of the regions surrounding the FMR1 and IDS genes have been determined (180 and 130 kb, respectively), along with an additional ~700 kb of the interval. This sequence has led to identification of the gene involved in FRAXE mental retardation. Additional sequence in Xq28 has been determined, including that of a cosmid containing the two genes, DXS1357E and a creatine transporter. This sequence has been duplicated to chromosome 16pl 1 in recent evolutionary history. Comparative sequence analysis reveals 94% sequence identity over 25 kb, and the presence of pentameric repeats which are likely to have mediated the duplication event. A number of technical advances in sequencing have been developed, including the use of BODIPY dyes in AB373 sequencing protocols, which has offered enhanced base calling due to reduced mobility shifting, improved single strand template protocols for much reduced cost, and streamlined informatics processes for assembly and annotation.

DOE Grant Nos. DE-FG05-92ER61401 and DE-FG03-94ER61830 and NIH Grant No. 5P30 HG00210.

*Sequence-Specific Proteins Binding to the Repetitive Sequences of High Eukaryotic Genome

Olga Podgornaya, Ivan Lobov, Ivan Matveev, Dmitry Lukjanov, Natella Enukashvily, and Elena Bugaeva
Institute of Cytology; Russian Academy of Sciences; St. Petersburg 194064, Russia
Telephone and Fax: +7-812/520-9703
podg@ivm.stud.pu.ru

Repetitive sequences occupy the most part of the whole eukaryotic genome but up to the last few years there has not been much interest in their role. The situation changed when alpha-satellites in human and minor satellites in mouse became candidates for centromere function responsibility. A number of centromere-specific proteins are under investigation but none seems to distinguish centromeric functions of exact sequences among long arrays of tandemly repeated satellites. The proteins associated with that array are poorly known. We are trying to find out what proteins are involved in maintaining the heterochromatin structure of different types of repetitive sequences.

The major proportion of total genomic satellite DNA remains attached to the nuclear matrix (NM) after DNase1 and high salt treatment. We followed this association in various steps during NM preparation by in situ hybridization with the mouse satellite probe. Two mouse species were used -*M. musculus* and *M. spretus*. Both contain the same repertoire of satellite DNAs but in different amounts. In *M. musculus* the centromeric heterochromatin contains major satellite (MA) as the principal component. In *M. spretus* the minor satellite (MI) is predominant. To test DNA-binding activity of the proteins after chromatography of the soluble NM proteins on cationic and anionic ion-exchange columns, gel shift assays were performed with cloned dimer of MA and a trimer of MI. To produce antibodies, the DNA-protein complexes obtained from large-scale gel-shift assays were isolated and injected into a guinea pig.

The gel shift assay with column fractions from *M. musculus* NM and MA shows a ladder of complexes. The complexes could be competed out with an excess of MA DNA but not with the same amount of *E. coli* DNA. Antibodies from the immune serum caused a hypershift of the MA/NM protein complexes. Preimmune serum at the same dilution did not alter the mobility of the complexes. A combination of western and Southern blots allows us to conclude that a protein with a molecular weight of about 80 kD and some similarity to the intermediate filaments is responsible for the MA/NM interaction.

Specific DNA-binding activity to the MI has been tested after column fractionation of the *M. spretus* NM extract. A ladder of complexes can be competed out with an excess of unlabeled MI but not *E. coli* or MA DNA. MI contains the CENPB-box sequence, which is the binding site for the protein CENPB, one of the centromeric proteins. Fractions from the NM extract with MI-specific binding activity do not contain CENPB, as shown by western blotting with anti-CENPB antibodies.

The same kind of work is going on with human analogs of MA and MI sequences, using large clones of satellite and alpha-satellite DNA and nuclear matrices.

There are few satellite DNA-binding proteins isolated, none of them directly from the NM. Our long-term aim is to understand the role of these proteins in heterochromatin formation and in heterochromatin association with NM.

Extracts from hand-isolated nuclear envelopes from frog oocytes were tested for the specific DNA-binding activity to (T2G4)116. A fragment of *Tetrahymena* telomere from a YAC plasmid was used as a labelled probe in a gel-shift assay. The DNA-protein complexes from the assay were cut out and injected into a guinea pig. The antibodies (AB) obtained stained one protein with an m.w. of about 70 kD in the nuclear envelope of the oocyte, nothing in the inner part of the oocyte, and 70 kD and 120 kD in the frog liver nuclei. The immunofluorescent AB stained fine patches on the oocyte nuclear envelope and a number of intranuclei spots in the frog blood cells.

The electron-microscope immuno-gold technique showed that the protein is localized in the outer surface of the oocyte nuclear envelope in cup-like structures. DNA-binding activity to the same sequence has been tested and found in the mouse nuclear matrix extracts. The activity could be eluted from the DEAE52 ion exchange column in 0.15 NaCl. The activity could be competed out with the fragment itself but not with *E. coli* DNA in the same amounts. AB stained a 70-kD protein in active fractions after ion exchange chromatography. In nuclear matrix preparations, the AB recognized a 120-kD protein as well. The AB caused hypershift of the complexes on the gel shift assay. The AB has some affinity to the keratins. In the mouse cell culture 3T3 line the staining is intranuclei, with fine dots forming chains surrounding dark areas, which do not correspond to the nucleoli.

Similar results were observed when a mouse cell line was transformed with head-and tail-less human keratin constructs (Bader et al., 1991, *J Cell Biol* **115**:1293). These results suggest that the nuclear proteins detected with the AB may be natural analogs of this artificial keratin construct. The pattern of staining did not resemble the picture of telomere-specific staining. Possibly the protein recognized intragenomic (T2G4)2 sequence, which is present in 25% of murine GenBank sequences rather than telomere. We are going to do immunocytochemical investigations of frog and mouse development in order to determine the point when transcription of the 120- kD protein is initiated and the staining becomes intranuclear.

As a continuation of the previous project the multiple alignment of all the *Alu* sequences from GenBank is going on. We are also trying to obtain antibodies to the main *Alu*-binding proteins to find out how many proteins could be bound to *Alu* sequence.

DOE Grant No. OR00033-93C1S014.

---

*Protein-Binding DNA Sequences

O.L. Polanovsky, A.G. Stepchenko, and N.N. Luchina
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
Fax: +7-095/135-1405, *pol@genome.eimb.rssi.ru*

POU domain of Oct-2 transcription factor binds octamer sequence ATGCAAAT and a number of degenerated sequences. It has been shown that POUs and POUh domains recognize left and right parts of the oct-sequence, respectively. The recognized sequences are partly overlapped in the native octamer. In the degenerated recognition sites these core sequences may be separated with a spacer up to four nucleotides. The obtained data changed our view on the number and structure of potential targets recognized on DNA by POU proteins.

Protein-DNA binding is realized due to interaction of a conservative amino acid residues with a DNA target. In POU proteins amino acid residues in positions 47 (Val), 50 (Cys) and 51 (Asn) of POUh domain are absolutely conservative. In order to examine a possible role of Val47 we substituted this residue by each of the 19 other amino acid residues and the interaction of the mutant proteins was investigated with homeospecific site and its variants (ATAANNN) and with oct sequence. It was shown that Ile47 mutant retains the affinity and specificity. Val replacement for Ser, Thr or His partially reduce the affinity.

Asn47 mutant sharply relax the specificity of protein-DNA recognition. Mutants at 47 position have much stronger effects on binding to homeospecific sites than to octamer motifs. Our data indicate that there is not a simple mono-letter code of protein/DNA recognition. It has been shown that this recognition is determined not only by the nature of the radicals involved in the contact but also by the structure of DNA binding domain as a whole and probably by cooperative interaction of POUs and POUh domains.

Proposals for 1997. The role of Cys50 in POU domain/DNA recognition will be investigated. This residue is absolutely conservative in POU proteins but it is variable in relative homeo-proteins. Our preliminary data allow to suppose that residue at position 50 of POU homeodomain have a key role in discrimination between TAAT-like and octamer sequences. The role of the nuleotides flanking DNA target will be investigated.

DOE Grant No. OR00033-93CIS005.

Relevant Publications

1. Stepchenko A.G. (1994) Noncanonical oct-sequences are targets for mouse Oct-2B transcription factor. FEBS Letters, V.337, P.175-178.
2. Stepchenko A.G., Polanovsky O.L. (1996) Interaction of Oct proteins with DNA. Molecular Biology, V.30, P.296-302.
3. Stepchenko A.G., Luchina N.N., Polanovsky O.L. The role of conservative Val47 for POU homeodomain/DNA recognition. FEBS Letters, in press.

Mapping

*Development of Intracellular Flow Karyotype Analysis

V.V. Zenin,[1] N.D. Aksenov,[1] A.N. Shatrova,[1] N.V. Klopov,[2]
L.S. Cram,[3] and **A.I. Poletaev**
Engelhardt Institute of Molecular Biology; Russian
Academy of Sciences; Moscow 117984, Russia
Poletaev: +7-095/135-9824, Fax: -1405
polet@polet.msk.su
[1]Institute of Cytology; Russian Academy of Sciences;
St. Petersburg, Russia
[2]St. Petersburg Institute of Nuclear Physics; Gatchina, Russia
[3]Los Alamos National Laboratory; Los Alamos, NM 87545

Instrumentation for univariate fluorescent flow analysis of chromosome sets has been developed for human cells. A new method of cell preparation and intracellular staining of chromosome with different dyes was developed and improved. Cells suspension for flow analysis must satisfy the following requirements: minimal amount of free chromosomes and debris (dead cells, cell fragments etc.); chromosomes structure must be stabilized inside mitotic cells; chromosomes must be stained inside the cells up to saturation with the used dyes; chromosomes must be able to release from cells with minimal possible mechanical treatment. The method includes enzyme treatment (chymotrypsin), incubation with saponin and separation of prestained cells from debris on sucrose gradient. The developed protocol was tested and improved in the course of several months of work and allows us to obtain a well stained sample with a minimal amount of contaminates [2].

A special magnetic mixing/stirring device was constructed to perform cell membrane breaking. It was placed inside the flow chamber of a serial flow cytometer ATC-3000 equipped with additional electronic card for time-gated data acquisition [1]. The rupturing of prestained mitotic cells is performed by means of a small magnetic rod vibrating in an alternative magnetic field. The efficiency of mitotic cells breaking with electromagnetic cell breaking device was tested using different human cell lines[2,3].

The device works in a stepwise mode: a defined volume of sample is delivered to the breaking chamber for rupturing mitotic cell (cells) for a defined time period, followed by buffer wash to move the released chromosomes from the breaking chamber to the point of the analysis. The information about the chromosomes appearing at the point of analysis is accumulated in list mode files, making it possible to resolve chromosome sets arising from single cells on the basis of time gating. The concentration of cells in the sample must be kept low to ensure that only one cell at a time enters the breaking device.

The developed software classifies chromosome sets according to different criteria: total number of chromosomes, overall DNA content in the set, and the number of chromosomes of certain type [2,3]. In addition it's possible to determine the presence of extra chromosomes or loss of chromosome types. Thus this approach combines the high performance of flow cytometry (quantitation and high throughput) with the advantages of image analysis (cell to cell karyotype analysis and skills of trained cytogeneticist). The data analysis capabilities offer extensive flexibility in determining important features of the karyotypes under study. This development offers the potential to duplicate most of what is determined by clinical cytogeneticists. The results now obtained are in good accordance with goals of the project formulated before [4].

DOE Grant No. OR00033-93CIS008.

References

[1]. V.V. Zenin, N.D. Aksenov, A.N. Shatrova, Y.V. Kravatsky, A. Kuznetzova, L.S. Cram, A.I. Poletaev. "Time-gated human chromosome flow analysis" XVII Congress of the International Society for Analytical Cytology, 1994, Lake Placid, USA, Cytometry Supplement 7, p. 68.

[2]. V.V. Zenin, N.D. Aksenov, A.N. Shatrova, Y.V. Kravatsky, A. Kuznetsova, L.S. Cram , A.I. Poletaev: "Time-gated flow analysis of human chromosomes"; DOE Human Genome Program, Contractor-Grantee Workshop IV, November 13-17, 1994; Santa Fe, New Mexico, p. 13.

[3]. V.V. Zenin, N.D. Aksenov, A.N. Shatrova, N.V. Klopov , L.S. Cram, A.I. Poletaev: "Cell by cell flow analysis of human chromosome sets"; DOE Human Genome Program, Contractor-Grantee Workshop V, January 28-February 1,1996; Santa Fe, New Mexico, p. 112.

[4]. Andrei I. Poletaev, Sergei I. Stepanov, Valeri V. Zenin, Nikolay Aksenov, Tatijana V. Nasedkina and Yuri V. Kravazky: "Development of Intracellular FlowKaryotype Analysis"; DOE Human Genome, 1993 Program Report, p.34-35.

Mapping and Sequencing with BACs and Fosmids

Ung-Jin Kim, Hiroaki Shizuya, and **Melvin I. Simon**
Division of Biology; California Institute of Technology;
Pasadena, CA 91125
Kim: 818/395-4901, Fax: /796-7066, *ung@caltech.edu*
Simon: 818/395-3944, Fax: /796-7066
simonm@starbase1.caltech.edu
http://www.tree.caltech.edu

BACs and fosmids are stable, nonchimeric, and highly representative cloning systems. BACs maintain large-fragment genomic inserts (100 to 300 kb) that are easily prepared for most types of experiments, including DNA sequencing.

We have improved the methods for generating BACs and developed extensive BAC libraries. We have constructed human BAC libraries with more than 175,000 clones from male fibroblast and sperm, and a mouse BAC library with more than 200,000 clones. We are currently expanding human library with the aim of achieving total 50X coverage human genomic library using sperm samples from anonymous donors.

The BAC libraries provide resources to bridge the gap between genetic-cytogenetic information and detailed physical characteristics of genomic regions that include DNA sequence information. They also provide reliable tools for generating a high-resolution, integrated map on which a variety of information and resources are correlated. Using primarily the human BAC library constructed from fibroblasts, we have assembled a physical contig map of chromosome 22 [1]. First, the entire library was screened by most of the known chromosome 22-specific markers that include cDNA, anonymous STS markers, FISH-mapped cosmids and fosmids, YAC-Alu PCR products, FISH-mapped BACs, and flow-sorted chromosome 22 DNA. The positive clones have been assembled into contigs by means of the STS-contents or other markers assigned to BAC clones. Most of the contigs were confirmed by using a restriction fingerprinting scheme originally developed by Sulston and Coulson, and modified in our laboratory. Currently, the contigs cover over 80% of the chromosome arm. Various physical or genetic landmarks on this chromosome can now be precisely localized simply by assigning them to BACs or contigs on the map. Using BAC end sequence information from each of the chromosome 22-specific BACs, it is now possible to close the gaps efficiently by screening deeper BAC libraries with new probes specific to the ends of contigs.

The resulting BAC contig map is now serving as a road map for sequencing the chromosome. Chromosome 22-specific BAC clones have been distributed to our collaborators including The Sanger Center and Dr. Bruce Roe in University of Oklahoma, and many of the clones have already been sequenced. BAC end sequencing scheme[2] will play a crucial role toward the complete sequencing of chromosome 22, and we are currently sequencing the ends of these BACs directly using the miniprepped BAC DNA as templates.

DOE Grant No. DE-FG03-89ER60891.

References

[1] Kim et al. (1996) A Bacterial Artificial Chromosome-based framework contig map of human chromosome 22q. Proc. Natl. Acad. Sci. USA v93 (13): pp6297-6301.
[2] Venter, C., Smith, H.O., and Hood, L. (1996) Nature 381: pp364-366.

Towards a Globally Integrated, Sequence-Ready BAC Map of the Human Genome

Ung-Jin Kim, Hiroaki Shizuya, and **Melvin I. Simon**
Division of Biology; California Institute of Technology; Pasadena, CA 91125
Kim: 818/395-4901, Fax: /796-7066, *ung@caltech.edu*
Simon: 818/395-3944, Fax: /796-7066
simonm@starbase1.caltech.edu
http://www.tree.caltech.edu

BAC clones are ideal for genome analysis since they are non-chimeric, stably maintain large fragment genomic inserts (100-300 kb)[1], and it is easy to prepare BAC DNA samples for most types of experiments including DNA sequencing[2]. We have improved BAC cloning technique in the past years and constructed >20X human BAC libraries. As BACs are proving to be the most efficient reagents for large scale genomic sequencing, we intend to increase the depth of the library to 50X genomic equivalence. Using the ESTs, especially the Unigenes that have been chromosomally assigned by other means such as Radiation Hybrid mapping and YAC-based STS content mapping, we plan to organize the BAC library into a mapped resource. The resulting BAC-EST framework map will provide a high resolution EST (or gene) map and instant entry points for gene finding and large scale genomic sequencing. We also intend to determine the end sequences of the BAC inserts from a significant number of the clones (at least 350,000 clones or 15X genomic equivalence) within two years [3]. All the BAC-EST mapping data and BAC end sequences will be made available via public databases and WEB servers. The mapping data and end sequence information will dramatically facilitate the process of finding clones that extend the sequenced regions with minimal overlaps. Thus, the tagged BAC libraries will serve as a reliable and facile sequence-ready resource and an organizing tool to support and coordinate simultaneously multiple sequencing projects all over the genome.

DOE Grant No. DE-FC03-96ER62242.

References

[1] Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M.I. (1992) Proc. Natl. Acad. Sci. USA 89, 8794-8797.
[2] Kim, U.-J., Birren, B.W., Yu-Ling Sheng, Tatiana Slepak, Valena Mancino, Cecilie Boysen, Hyung-Lyun Kang, Melvin I. Simon, and Hiroaki Shizuya. (1996) Genomics 34, 213-218.
[3] Venter, C, Smith, H.O., and Hood, L. (1996) Nature 381: pp364-366.

Generation of Normalized and Subtracted cDNA Libraries to Facilitate Gene Discovery

Marcelo Bento Soares, Maria de Fatima Bonaldo, Pierre Jelenc, and Susan Baumes
Department of Psychiatry; Columbia University; and The New York State Psychiatric Institute; New York, NY 10032
212/960-2313, Fax: /781-3577,
cuc@cuccfa.ccc.columbia.edu

Large-scale single-pass sequencing of cDNA clones randomly picked from libraries has proven quite powerful to identify genes and the use of normalized libraries in which the frequency of all cDNAs is within a narrow range has been shown to expedite the process by minimizing the redundant identification of the most prevalent mRNAs. In an attempt to contribute to the ongoing gene discovery efforts, we have further optimized our original procedure for construction of normalized directionally cloned cDNA libraries[1] and we have successfully applied it to generate a number of human cDNA libraries from a variety of adult and fetal tissues [2]. To date we have constructed libraries from infant brain, fetal brain, adult brain, fetal liver-spleen, full-term and 8-9 week placentae, adult breast, retina, ovary tumor, melanocytes, parathyroid tumor, senescent fibroblasts, pineal glands, multiple sclerosis plaques, testis, B cells, fetal heart, fetal lung, 8-9 week fetuses and pregnant uterus. Several additional libraries are currently in preparation. All libraries have been contributed to the IMAGE consortium, and they are being widely used for sequencing and mapping.

However, given the large scale nature of the ongoing sequencing efforts and the fact that a significant fraction of the human genes has been identified already, the discovery of novel cDNAs is becoming increasingly more challenging. In an effort to expedite this process further, in collaboration with Greg Lennon (LLNL) we have developed and applied subtractive hybridization strategies to eliminate pools of sequenced cDNAs from libraries yet to be surveyed. Briefly, single-stranded DNA obtained from pools of arrayed and sequence I.M.A.G.E. clones are used as templates for PCR amplification of cDNA inserts with flanking T7 and T3 primers. PCR amplification products are then used as drivers in hybridizations with normalized libraries in the form of single-stranded circles. The remaining single-stranded circles (subtracted library) are purified by hydroxyapatite chromatography, converted to double-stranded circles and electroporated into bacteria. Preliminary characterization of a subtracted fetal liver-spleen library indicates that the procedure is effective to enhance the representation of novel cDNAs.

In an effort to enhance the representation of full-length cDNAs in our libraries, as we strive towards our final objective of generating full-length normalized cDNA libraries, we have adapted our normalization protocol to take advantage of the fact that it is now possible to produce single-stranded circles in vitro by sequentially digesting supercoiled plasmids with Gene II protein and Exonuclease III (Life Technologies). This has proven significant because it circumvents the biases introduced by differential growth of clones containing small and large cDNA inserts when single-strands are produced in vivo upon superinfection with a helper phage.

DOE Grant No. DE-FG02-91ER61233.

References

[1] Soares, M.B., Bonaldo, M.F., Su, L., Lawton, L. & Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. USA 91(20), 9228-9232.

[2] Bonaldo, M.F., Lennon, G. and Soares, M.B. (1996). Normalization and subtraction: Two approaches to facilitate gene discovery. Genome Research 6, 791-806.

Mapping in Man-Mouse Homology Regions

Lisa Stubbs, Johannah Doyle, Ethan Carver,
Mark Shannon, Joomyeong Kim, Linda Ashworth,[1] and
Elbert Branscomb[1]
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831
423/574-0854, Fax: -1283, *stubbsl@bioax1.bio.ornl.gov* or *stubbslj@ornl.gov*
[1]Human Genome Center; Lawrence Livermore National Laboratory; Livermore, CA 94550

Numerous studies have confirmed the notion that mouse and human chromosomes resemble each other closely within blocks of syntenic homology that vary widely in size, containing from just a few to several hundred related genes. Within the best-mapped of these homologous regions, the presence and location of specific genes can be accurately predicted in one species, based upon the mapping results obtained in the other. In addition, information regarding gene function derived from the analysis of human hereditary traits or mapped murine mutations, can also be extrapolated from one species to another. However, syntenic relationships are still not established for many human regions, and local rearrangements including apparent deletions, inversions, insertions, and transposition events, complicate most of the syntenically homologous regions that appear simple on the gross genetic level. Because of these complications, the power of prediction afforded in any homology region increases tremendously with the level of resolution and degree of internal consistency associated with a particular set of comparative mapping data. Our groups have been interested in further defining the borders of syntenic linkage groups in human and mouse, upon elucidating mechanisms behind evolutionary rearrangements that distinguish chromosomes of mammalian species, and upon devising means of exploiting the relationships between the two genomes for the discovery and analysis of new genes and other functional units in mouse and man.

One of the larger contiguous blocks of mouse-human genomic homology includes the proximal portion of mouse chromosome 7 (Mmu7). Detailed analysis of this large region of mouse-human homology have served as the initial focus of these collaborative studies. Our results have shown that gene content, order and spacing are remarkably well-conserved throughout the length of this approximately 23 cM/29 Mb region of mouse-human homology, except for six internal rearrangements of gene sequence in mouse relative to man. One of these differences involve a small segment of H19ql3.4 genes whose murine counterparts have been transposed out of the large Mmu7/H19q conserved synteny region into a separate linkage group located on mouse chromosome 17. The six internal rearrangements, including two transpositions and four local inversions, are clustered together at two sites; our data suggest that the rearrangements occurred in a coincident fashion, or were commonly associated with unstable DNA sequences at those sites. Interestingly, both rearranged regions are occupied by large tandemly clustered gene families, suggesting that these locally repeated sequences may have contributed to their evolutionary instability. The structure and conserved functions of genes within these and other clustered gene families located on H19 also represent an active line of interest to our group. More recently, we have extended mapping studies to include clustered gene families located in other chromosomal regions, and are working to define the borders of mouse-human syntenic segments on a broader, genome-wide scale.

DOE Contract No. DE-AC05-96OR22464 and Contract No. W-7405-ENG-48 with Lawrence Livermore National Laboratory.

Positional Cloning of Murine Genes

Lisa Stubbs, Cymbeline Culiat, Ethan Carver, Johannah Doyle, Laura Chittenden, Mitchell Walkowicz, Nestor Cacheiro, Greg Lennon,[1] Gary Wright,[2] Joe Rutledge,[3] Robert Nicholls,[4] and Walderico Generoso
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
423/574-0854, Fax: -1283, *stubbsl@bioax1.bio.ornl.gov* or *stubbslj@ornl.gov*
[1]Human Genome Center; Lawrence Livermore National Laboratory; Livermore, CA 94550
[2]University of Texas Southwestern Medical Center at Dallas; Dallas, TX 75235
[3]Children's Hospital and Medical Center; University of Washington School of Medicine; Seattle, WA 98105
[4]Department of Genetics; Case Western Reserve University; Cleveland, Ohio

Chromosome rearrangements, notably deletions and translocations, have proved invaluable as tools in the mapping and molecular cloning of a acquired and inherited human diseases. Because balanced translocations are cytologically visible, and generally produce profound disturbances in both gene expression and DNA structure without necessarily disturbing the structure of multiple genes, this type of mutation provides an especially valuable "tag" that greatly simplifies mapping, cloning, and assessment of candidate genes associated with a disease. Although balanced translocations are relatively rare in human populations, they are readily induced in the mouse. Using various mutagenesis protocols, we have generated numerous translocation-bearing mutant mouse strains that display an impressive variety of health-related anomalies, including obesity, polycystic kidneys, gastrointestinal disorders, limb and skeletal deformities, neural tube defects, ataxias, tremors, hereditary deafness and blindness, reproductive dysfunction, and complex behavioral defects. The ability to map the genes associated with translocation breakpoints cytogenetically, first crudely through straightforward banding techniques and then to a higher level of resolution using fluorescence in situ hybridization methods, allows us to avoid the costly and time-consuming crosses that are required for the mapping of most mutant genes. With this rapidly-obtained, crude-level mapping information available, we can readily assess possible relationships between newly arising mutant phenotypes and linked candidate genes or related diseases that map to homologous regions of the human genome. Using this approach, we have recently begun to define the map positions of several mutations. Mapping results have led us to the identification of candidate genes for two mutations: one associated with congenital deafness and predisposition to severe gastric ulcers, and another associated with late-onset obesity. So far, we have characterized only a fraction of the mouse strains that comprise this valuable, recently-generated mutant collection in detail. As a integral part of this program, we are actively exploring new strategies and integrating information, technology and resources derived from the Human Genome research effort, that promise to increase the efficiency of breakpoint mapping and cloning dramatically. The mutations are scattered widely throughout the mouse genome corresponding to a broad selection of human homology regions. As new breakpoints are mapped, and large numbers of newly-sequenced cDNA clones are assigned to the mouse and human maps, the potential for rapid association between cloned gene and mapped mutation will increase dramatically. This large collection of murine translocation mutants therefore represents a powerful resource for linking mapped cDNA clones to health-related phenotypes throughout the genome.

In addition to the analysis of translocation mutants, we have also characterized other types of mouse mutations, including: (1) *tottering* and *leaner*, allelic mutations associated with ataxia and epilepsy in mice, and representing murine models for human diseases, familial hemiplaegic migraine and episodic ataxia, respectively; and (2) *jdf2*, a locus associated with mutations causing runting, neuromuscular tremors and male sterility which is located in a mouse region related to the Prader Willi-Angleman syndrome gene interval of human 15q11-q13. Both sets of mutations affect large, complex, and highly conserved genes, and provide important animal models for the exploration of the diverse roles their human counterparts may play in human disease. In concert with these gene cloning studies, we have been involved in exploring new means of exploiting mouse-human genomic conservation in the isolation of functionally-significant sequences from large cloned regions of human DNA. The methods we have developed hold great promise as an efficient tool for gene discovery in cloned genomic regions.

DOE Contract No. DE-AC05-96OR22464.

Mapping

Human Artificial Episomal Chromosomes (HAECS) for Building Large Genomic Libraries

Min Wang, Panayotis A. Ioannou,[2] Michael Grosz, Subrata Banerjee, Evy Bashiardes,[2] Michelle Rider, Tian-Qiang Sun,[1] and **Jean-Michel H. Vos**[1]
Lineberger Comprehensive Cancer Center and [1]Department of Biochemistry and Biophysics; University of North Carolina; Chapel Hill, NC 27599
Vos: 919/966-3036, Fax: -3015, *vos@med.unc.edu*
[2]The Cyprus Institute of Neurology and Genetics; Nicosia, Cyprus

Of some 100,000 human genes, only a few thousand have been cloned, mapped or sequenced so far. Much less is known about other chromosomal regions such as those involved in DNA replication, chromatin packaging, and chromosome segregation. Construction of detailed physical maps is only the first step in localizing, identifying and determining the function of genetic units in human cells. Studying human gene function and regulation of other critical genomic regions that span hundreds of kilobase pairs of DNA requires the ability to clone an entire functional unit as a single DNA fragment and transfer it stably into human cells.

We have developed a human artificial episomal chromosome (HAEC) system based on latent replication origin of the large herpes Epstein-Barr virus (EBV) for the propagation and stable maintenance of DNA as circular minichromosomes in human cells.[1,2] Individual HAECS carried human genomic inserts ranging from 60 to 330 kb and appeared genetically stable. An HAEC library of 1500 independent clones carrying random human genomic fragments with average sizes of 150 to 200 kb was established and allowed recovery of the HAEC DNA. This autologous HAEC system with human DNA segments directly cloned in human cells provides an important tool for functional study of large mammalian DNA regions and gene therapy.[3,4]

Current efforts are focused on (a) shuttling large BAC/PAC genomic inserts in human and rodent cells and (b) packaging BAC/PAC/HAEC clones as large infectious Herpes Viruses for shuttling genomic inserts between mammalian cells and (c) constructing bacterial-based human and rodent HAEC libraries. (a) We have designed a "pop-in" vector, which can be inserted into current BAC-or PAC-based clone via site-specific integration. This "CRE-LOXP"-mediated system has been used to establish BAC/PAC up to 250 kb in size in human cells as HAECS. (b) We have obtained packaging of 160-180 kb exogenous DNA into infectious virions using the human lymphotropic Epstein-Barr virus. After delivery into human beta-lymphoblasts cells the HAEC DNA was stably established as 160-180 kb functional autonomously replicating episomes.[5,7] We have also generated a hybrid BAC/HAEC vector, which can shuttle large DNA inserts, i.e., at least up to 260 kb, between bacteria and human cells. Such a system is being used to develop large insert libraries, whose clones can be directly transferred into human or rodent cells for functional analysis. These HAEC-derived systems will provide useful molecular tools to study large genetic units in humans and rodents, and complement the functional interpretation of current sequencing efforts.

DOE Contract No. DE-FG05-91ER61135.

References

[1] Sun, T.-Q., Fenstermacher, D. & Vos, J.-M.H. Human artificial episomal chromosomes for cloning large DNA in human cells. Nature Genet 8, 33-41 (1994).
[2] Sun, T.-Q. & Vos, J.-M.H. Engineering of 100-300 kb of DNA as persisting extrachromosomal elements in human cells using the HAEC system in Methods molec. Genet. (ed. Adolph, K.W.) (Academic Press, San Diego, CA, 1995).
[3] Vos, J.-M.H. Herpes viruses as Genetic Vectors in Viruses in Human Gene Therapy (ed. Vos, J.-M.H.) 109-140 (Carolina Academic Press & Chapman & Hall, Durham N.C., USA & London, UK, 1995).
[4] Kelleher, Z. & Vos, J.-M. Long-Term Episomal Gene Delivery in Human Lymphoid Cells using Human and Avian Adenoviral-assisted Transfection. Biotechniques 17, 1110-1117 (1994).
[5] Banerjee, S., Livanos, E. & Vos, J.-M.H. Therapeutic Gene Delivery in Human beta-lymphocytes with Engineered Epstein-Barr Virus. Nature Medicine 1, 1303–1308 (1995).
[6] Sun, T.-Q., Livanos, E., & Vos, J.-M.H. Engineering a mini-herpesvirus as a general strategy to transduce up to 180 kb of functional self-replicating human mini-chromosomes. Gene Therapy 3, 1081–1088 (1996).
[7] Wang, S. & Vos, J.-M.H. An HSV/EBV based vector for High Efficient Gene Transfer to Human Cells in vitro/in vivo. J. Virol. 70, 8422–8430 (1996).

*Cosmid and cDNA Map of a Human Chromosome 13q14 Region Frequently Lost at B Cell Chronic Lymphocytic Leukemia

N.K. Yankovsky, B.I. Kapanadze, A.B. Semov, A.V. Baranova, and G.E. Sulimova
N.I. Vavilov Institute of General Genetics; Moscow 117809, Russia
+7-095/135-5363, Fax: -1289, *yankovsky@vigg.ru* and *bion@glas.apc.org* (send to both addresses)

We are mapping a human chromosome 13q14 region frequently lost at human blood malignancy cold B cell chronic lymphocytic leukemia (BCLL). The final goal of the project is to find putative oncosupressor gene lost in the region at BCLL. We have constructed a cosmid contig between D13S1168 and D13S25 loci in the region. The interval had been shown to be in the center of the BCLL associated deletions. The contig consists of more than 100 cosmids from LANL human chromosome 13 specific library (LA13NC01). We estimated the distance between D13S1168 and D13S25 loci as about 540 kb. We are constructing a transcriptional map of the region. Seven different cDNA clones were found with two of the cosmid clones. All cosmids corresponding to the minimal tilling path between D13S1168 and D13S25 are being used as probes for screening new cDNA clones. I.M.A.G.E. Consortium (LLNL) cDNA clones assigned to 13q14 will be mapped against the cosmid contig. Mapped cDNA clones will be checked as candidate oncosupressor genes for BCLL.

BCM Server Core

Daniel Davison and Randall Smith
Baylor College of Medicine; Houston, TX 77030
713/798-3738, Fax: -3759, *davison@bcm.tmc.edu*
http://www.bcm.tmc.edu

We are providing a variety of molecular biology-related search and analysis services to Genome Program investigators to improve the identification of new genes and their functions. These services are available via the BCM Search Launcher World Wide Web (WWW) pages which are organized by function and provide a single point-of-entry for related searches. Pages are included for 1) protein sequence searches, 2) nucleic acid sequence searches, 3) multiple sequence alignments, 4) pairwise sequence alignments, 5) gene feature searches, 6) sequence utilities, and 7) protein secondary structure prediction. The Protein Sequence Search Page, for example, provides a single form for submitting sequences to WWW servers that provide remote access to a variety of different protein sequence search tools, including BLAST, FASTA, Smith-Waterman, BEAUTY, BLASTPAT, FASTAPAT, PROSITE, and BLOCKS searches. The BCM Search Launcher extends the functionality of other WWW services by adding additional hypertext links to results returned by remote servers. For example, links to the NCBI's Entrez database and to the Sequence Retrieval System (SRS) are added to search results returned by the NCBI's WWW BLAST server. These links provide easy access to Medline abstracts, links to related sequences, and additional information which can be extremely helpful when analyzing database search results. For novice or infrequent users of sequence database search tools, we have pre-set the parameter values to provide the most informative first-pass sequence analysis possible.

A batch client interface to the BCM Search Launcher for Unix and Macintosh computers has also been developed to allow multiple input sequences to be automatically searched as a background task, with the results returned as individual HTML documents directly on the user's system. The BCM Search Launcher as well as the batch client are available on the WWW at URL http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html.

The BCM/UH Server Core provides the necessary computational resources and continuing support infrastructure for the BCM Search Launcher. The BCM/UH Server Core is composed of three network servers and currently supports electronic mail and WWW-based access; ultimately, specialized client-server access will also be provided. The hardware used includes a 2048-processor MasPar massively parallel MIMD computer, a DEC Alpha AXP/OSF1, a Sun 2-processor SparcCenter 1000 server, and several Sun Sparc workstations.

In addition to grouping services available elsewhere on the WWW and providing access to services developed at BCM and UH, the BCM/UH Server Core will also provide access to services from developers who are unwilling or unable to provide their own Internet network servers.

Grant Nos.: DOE, DE-FG03-9SER62097/A000; National Library of Medicine, R01-LM05792; National Science Foundation, BIR 91-11695; National Research Service Award, F32-HG00133-01; NIH, P30-HG00210 and R01-HG00973-01.

A Freely Sharable Database-Management System Designed for Use in Component-Based, Modular Genome Informatics Systems[†]

Steve Rozen,[1] Lincoln Stein,[1] and **Nathan Goodman**
The Jackson Laboratory; Bar Harbor, ME 04609
Goodman: 207/288-6158, Fax: -6078, *nat@jax.org*
[1]Whitehead Institute for Biomedical Research; Cambridge, MA 02139
http://goodman.jax.org
http://www.genome.wi.mit.edu/informatics/workflow

We are constructing a data-management component, built on top of commercial data-management products, tuned to the requirements of genome applications. The core of this genome data manager is designed to:

- support the semantic and object-oriented data models that have been widely embraced for representing genome data,
- provide domain-specific built-in types and operations for storing and querying bimolecular sequences,
- provide built-in support for tracking laboratory work flows, and admit further extensions for other special-purpose types,
- allow core facilities to be readily extended to meet the diverse needs of biological applications

The core data manager is being constructed on top of Sybase, Oracle, and Informix Universal Server. The software is available free of charge and is freely redistributable.

We will be reporting progress on the core data manager's architecture and interface at the URLs above, and we solicit comments on its design.

DOE Grant No. DE-FG02-95ER62101.

[†] Originally called Database Management Research for the Human Genome Project, this project was initiated in 1995 at the Massachusetts Institute of Technology–Whitehead Institute.

---

*Projects designated by an asterisk received small emergency grants following December 1992 site reviews by David Galas (formerly DOE Office of Health and Environmental Research, which was renamed Office of Biological and Environmental Research in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (Lawrence Livermore National Laboratory).

A Software Environment for Large-Scale Sequencing

Mark Graves
Department of Cell Biology; Baylor College of Medicine; Houston, TX 77030
713/798-8271, Fax: -3759; *mgraves@bcm.tmc.edu*
http://www.bcm.tmc.edu
http://stork.bcm.tmc.edu/gfp

Our approach is to implement software systems which manage primary laboratory sequence data and explore and annotate functional information in genome sequence and gene products.

Three software systems have been developed and are being used: two sequence data managers which use different sequence assembly packages, FAK and Phrap, and a series of analysis and annotation tools which are available via the Internet. In addition, we have developed a prototype application for data mining of sequence data as it is related to metabolic pathways.

Products of this project are the following:

1. GRM -a sequence reconstruction manager using the FAQ assembly engine (available since October 1995).

2. GFP -a sequence finishing support tool using the Phrap assembly engine (available since March 1996).

3. A series of gene recognition tools (available since early 1996).

4. A tool for visualizing metabolic pathways data and exploring sequence data related to metabolic pathways (prototype available since August 1996).

DOE Grant No. DE-FG03-94ER61618.

Generalized Hidden Markov Models for Genomic Sequence Analysis

David Haussler, Kevin Karplus,[1] and Richard Hughey[1]
Computer Science Department and [1]Computer Engineering Department; University of California; Santa Cruz, CA 95064
408/459 2105, Fax: -4829, *haussler@cse.ucsc.edu*
http://www.cse.ucsc.edu/research/compbio
http://www-hgc.lbl.gov/projects/genie.html

We have developed an integrated probabilistic method for locating genes in human DNA based on a generalized hidden Markov model (HMM). Each state of a generalized HMM represents a particular kind of region in DNA, such as an initial exon for a gene. The states are connected by transitions that model sites in DNA between adjacent regions, e.g. splice sites. In the full HMM, parametric statistical models are estimated for each of the states and transitions. Generalized HMMs allow a variety of choices for these models, such as neural networks, high order Markov models, etc. All that is required is that each model return a likelihood for the kind of region or transition it is supposed to model. These likelihoods are then combined by a dynamic programming method to compute the most likely annotation for a given DNA contig. Here the annotation simply consists of the locations of the transitions identified in the DNA, and the labeling of the regions between transitions with their corresponding states.

This method has been implemented in the genefinding program Genie, in collaboration with Frank Eeckman, Martin Reese and Nomi Harris at Lawrence Berkeley Labs. David Kulp, at UCSC, has been responsible for the core implementation. Martin Reese developed the splice site models, promoter models, and datasets. You can access Genie at the second www address given above, submit sequences, and have them annotated. Nomi Harris has written a display tool called Genotater that displays Genie's annotation along with the annotation of other genefinders, as well as the location of repetitive DNA, BLAST hits to the protein database, and other useful information. Papers and further information about Genie can be found at the first www address above. Since the ISMB '96 paper, Genie's exon models have been extended to explicitly incorporate BLAST and BLOCKS database hits into their probabilistic framework. This results in a substantial increase in gene predicting accuracy. Experimental results in tests using a standard set of annotated genes showed that Genie identified 95% of coding nucleotides correctly with a specificity of 88%, and 76% of exons were identified exactly.

DOE Grant No. DE-FG03-95ER62112.

Identification, Organization, and Analysis of Mammalian Repetitive DNA Information

Jerzy Jurka
Genetic Information Research Institute; Palo Alto, CA 94306
415/326-5588 Fax: -2001, *jurka@gnomic.stanford.edu*
http://charon.lpi.org

There are three major objectives in this project: organization of databases of mammalian repetitive sequences, development of specialized software for analysis of repetitive DNA, and sequence studies of new mammalian repeats.

Our approach is based on extensive usage of computer tools to investigate and organize publicly available sequence information. We also pursue collaborative research with experimental laboratories. The results are widely disseminated via the internet, peer reviewed scientific publications and personal interactions. Our most recent research concentrates on mechanisms of retroposon integration in mammals (Jurka, J., PNAS, in press; Jurka, J and Klonowski, P., J. Mol. Evol. 43:685-689).

We continue to develop reference collections of mammalian repeats which became a worldwide resource for annotation and study of newly sequenced DNA. The reference collections are being revised annually as part of a larger database of repetitive DNA, called Repbase. The recent influx of sequence data to public databases created an unprecedented need for automatic annotation of known repetitive elements. We have designed and implemented a program for identification and elimination of repetitive DNA known as CENSOR.

Reference collections of mammalian repeats and the CENSOR program are available electronically (via anonymous ftp to ncbi.nih.gov; directory repository/repbase). CENSOR can also be run via electronic mail (mail "help" message to censor@charon.lpi.org).

DOE Grant No. DE-FG03-95ER62139.

*TRRD, GERD and COMPEL: Databases on Gene-Expression Regulation as a Tool for Analysis of Functional Genomic Sequences

A.E. Kel, O.A. Podkolodnaya, O.V. Kel, A.G. Romaschenko, E. Wingender,[1] G.C. Overton,[2] and **N.A. Kolchanov**
Institute of Cytology and Genetics; Novosibirsk, Russia
Kolchanov: +7-3832/353-335, Fax: -336 or /356-558,
kol@benpc.bionet.nsc.ru
http://transfac.gbf-braunschweig.de
[1]Gesellschaft für Biotechnologische Forschung; Braunschweig, Germany
[2]Department of Genetics; University of Pennsylvania School of Medicine; Philadelphia, PA 19104-6145

The database on transcription regulatory regions in eukaryotic genomes (TRRD) has been developed [1] (http://www.bionet.nsk.su/TRRD.html; ftp://ftp.bionet.nsk.su/pub/trrd/). The main principle of data representation in TRRD is modular structure and hierarchy of transcription regulatory regions. TRRD entry corresponds to a gene as entire unit. Information on gene regulation is provided (cell-cycle and cell type specificity, developmental stage-specificity, influence of various molecular signals on gene expression). TRRD database contains information about structural organization of gene transcription regulatory region. TRRD contains description of known promoters and enhancers in 5', 3' regions and in introns. Description of binding sites for transcription factors includes nucleotide sequence and precise location, name of factors that bind to the site, experimental evidences for the binding site revealing. We provide cross-references to TRANSFAC database [2] for both sites and factors as well as for genes. TRRD 3.3 release includes 340 vertebrate genes.

The Gene Expression Regulation Database (GERD) collects information on features of genes expression as well as information about gene transcription regulation. The current release of GERD contains 75 entries with information on expression regulation of genes expressed in hematopoietic tissues in the course of ontogenesis and blood cells differentiation. COMPEL database contains information about composite elements which are functional units essential for highly specific transcription regulation [3]. Direct interactions between transcription factors binding to their target sites within composite elements result in convergence of different signal transduction pathways. Nucleotide sequences and positions of composite elements, binding factors and types of their DNA binding domains, experimental evidence confirming synergistic or antagonistic action of factors are registered in COMPEL. Cross-references to TRANSFAC factors table are given. TRRD and COMPEL are provided by cross-references to each other. COMPEL 2.1 release includes 140 composite elements.

We have developed a software for analysis of transcription regulatory region structure. The CompSearch program is based on oligonucleotide weight matrix method. To collect sets of binding sites for the matrixes construction we have used TRANSFAC and TRRD databases. The CompSearch program takes into account the fine structure of experimentally confirmed NFATp/AP-1 composite elements collected in COMPEL (distances between binding sites in composite elements, their mutual orientation). By means of the program we have found potential composite elements of NFATp/AP-1 type in the regulatory regions of various cytokine genes. Analysis of composite elements could be the first approach to reveal specific patterns of transcription signals encoding regulatory potential of eukaryotic promoters.

References

1. Kel O.V., Romaschenko A.G., Kel A.E., Naumochkin A.N., Kolchanov N.A. Proceedings of the 28th Annual Hawaii International Conference on System Sciences [HICSS]. (1995), v.5, Biotechnology Computing, IEE Computer Society Press, Los Alamos, California, p. 42-51.
2. Wingender E., Dietze P., Karas H., and Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites (1996). Nucl. Acids Res., 1996, v. 24, pp. 238-241.
3. Kel O.V., A.G. Romaschenko, A.E. Kel, E. Wingender, N.A. Kolchanov. A compilation of composite regulatory elements affecting gene transcription in vertebrates (1995). Nucl. Acids Res., v. 23, pp. 4097-4103.

(abstract continued)

Recent Publications

Kel, A., Kel, O., Ischenko, I., Kolchanov, N., Karas, H., Wingender, E. and Sklenar, H. (1996). TRRD and COMPEL databases on transcription linked to TRANSFAC as tools for analysis and recognition of regulatory sequences. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB'96), R. Hofestadt, T. Lengauer, M. Löffler, D. Schomburg (eds.). University of Leipzig, Leipzig 1996, pp. 113-117.

Wingender, E., Kel, A. E., Kel, O. V., Karas, H., Heinemeyer, T., Dietze, P., Knueppel, R., Romaschenko, A. G. and Kolchanov, N. A. (1997). TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation. Nucleic Acids Res., in press.

Ananko E.A., Ignatieva E.V., Kel A.E., Kolchanov N.A (1996). WWWTRRD: Hypertext information system on transcription regulation. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB'96), R. Hofestadt, T. Lengauer, M. Löffler, D. Schomburg (eds.). University of Leipzig, Leipzig 1996, pp. 153-155.

A.E. Kel, O.V. Kel, O.V. Vishnevsky, M.P. Ponomarenko, I.V. Ischenko, H. Karas, N.A. Kolchanov, H. Sklenar, E. Wingender (1997). TRRD and COMPEL databases on transcription linked to TRANSFAC as tools for analysis and recognition of regulatory sequences. (1997) LECTURE NOTES IN COMPUTER SCIENCE, in press.

Holger Karas, Alexander Kel, Olga Kel, Nikolay Kolchanov, and Edgar Wingender (1997). Integrating knowledge on gene regulation by a federated database approach: TRANSFAC, TRRD and COMPEL. Jurnal Molekularnoy Biologii (Russian), in press.

Kel A.E., Kolchanov N.A., Kel O.V., Romaschenko A.G., Ananko E.A., Ignatyeva E.V., Merkulova T.I., Podkolodnaya O.A., Stepanenko I.L., Kochetov A.V., Kolpakov F.A., Podkolodniy N.L., Naumochkin A.A. (1997). TRRD: A database on transcription regulatory regions of eukaryotic genes. Jurnal Molekularnoy Biologii (Russian) in press.

O.V. Kel, A.E. Kel, A.G. Romaschenko, E. Wingender, N.A. Kolchanov (1997). Composite regulatory elements: classification and description in the COMPEL data base. Jurnal Molekularnoy Biologii (Russian), in press.

Data-Management Tools for Genomic Databases

Victor M. Markowitz and **I-Min A. Chen**
Information and Computing Sciences Division; Lawrence Berkeley National Laboratory; Berkeley, CA 94720
510/486-6835, Fax: -4004, *vmmarkowitz@lbl.gov*
http://gizmo.lbl.gov/opm.html

The Object-Protocol Model (OPM) data management tools provide facilities for constructing, maintaining, and exploring efficiently molecular biology databases. Molecular biology data are currently maintained in numerous molecular biology databases (MBDs), including large archival MBDs such as the Genome Database (GDB) at Johns Hopkins School of Medicine, the Genome Sequence Data Base (GSDB) at the National Center for Genome Resources, and the Protein Data Bank (PDB) at Brookhaven National Laboratory. Constructing, maintaining, and exploring MBDs entail complex and time-consuming processes.

The goal of the Object-Protocol Model (OPM) data management tools is to provide facilities for efficiently constructing, maintaining, and exploring MBDs, using application-specific constructs on top of commercial database management systems (DBMSs). The OPM tools will also provide facilities for reorganizing MBDs and for exploring seamlessly heterogenous MBDs. The OPM tools and documentation are available on the Web and are developed in close collaboration with groups maintaining MBDs, such as GDB, GSDB, and PDB.

Current work focuses on providing new facilities for constructing and exploring MBDs. The specific aims of this work are:

(1) Extend the OPM query language with additional constructs for expressing complex conditions, and enhance the OPM query optimizer for generating more efficient query plans.

(2) Develop enhanced OPM query interfaces supporting MBD-specific data types (e.g., protein data type) and operations (e.g., protein data display and 3D search), and assisting users in specifying and interpreting query results.

(3) Provide support for customizing MBD interfaces.

(4) Extend the OPM tools with facilities for managing permissions (object ownership) in MBDs, and for physical database design of relational MBDs, including specification of indexes, allocation of segments, and handling of redundant (denormalized) data.

(5) Develop OPM tools for constructing and maintaining multiple OPM views for both relational and non-relational (e.g., ASN.1, AceDB) MBDs. For a given MBD, these tools will allow customizing different OPM views for different groups of scientists. For heterogeneous MBDs, this tool will allow exploring them using common OPM interfaces.

(6) Develop tools for constructing OPM based multidatabase systems of heterogeneous MBDs and for exploring and manipulating data in these MBDs via OPM interfaces. As part of this effort, the OPM-based multidatabase system which consists currently of GDB 6.0 and GSDB 2.0, will be extended to include additional MBDs, primarily GSDB 2.2 (when it becomes available), PDB, and Genbank.

(7) Develop facilities for reorganizing OPM-based MBDs.The database reorganization tools will support automatic generation of procedures for reorganizing MBDs following restructuring (revision) of MBD schemas.

In the past year, the OPM data management tools have been extended in order to address specific requirements of developing MBDs such as GDB 6 and the new version of PDB.

The current version of the OPM data management tools (4.1) was released in June 1996 for Sun/OS, Sun/Solaris and SGI. The following OPM tools are available on the Web at http://gizmo.lbl.gov/opm.html:

(1) an editor for specifying OPM schemas;

(2) a translator of OPM schemas into relational database specifications and procedures;

(3) utilities for publishing OPM schemas in text (Latex), diagram (Postscript), and Html formats;

(4) a translator of OPM queries into SQL queries;

(5) a retrofitting tool for constructing OPM schemas (views) for existing relational genomic databases;

(6) a tool for constructing Web-based form interfaces to MBDs that have an OPM schema; this tool was developed by Stan Letovsky at Johns Hopkins School of Medicine, as part of a collaboration.

The OPM data management tools have been highly successful in developing new genomic databases, such as GDB 6 (released in January 1996; http://gdbgeneral.gdb.org/gdb/) and the relational version of PDB (http://terminator.pdb.bnl.gov:4148), and in constructing OPM views and interfaces for existing genomic databases such as GSDB 2.0. The OPM data management tools are currently used by over ten groups in USA and Europe. The research underlying these tools is described in several papers published in scientific journals and presented at database and genome conferences.

In the past year the OPM tools have been presented at database and bioinformatics conferences, including the International Symposium on Theoretical and Computational Genome Research, Heidelberg, Germany, March 1996, the Workshop on Structuring Biological Information, Heidelberg, Germany, March 1996, the Meeting on Genome Mapping and Sequencing, Cold Spring Harbor, May 1996, the International Sybase User Group Conference, May 1996, the Bioinformatics -Structure Conference, Jerusalem, November 1996, and the Pacific Symposium on Bioinformatics, January 1997.

The results of the research and development underlying the OPM tools work have been presented in papers published in proceedings of database and bioinformatics conferences; these papers are available at http://gizmo.lbl.gov/opm.html#Publications.

DOE Contract No. DE-AC03-76SF00098.

The Genome Topographer: System Design

S. Cozza, D. Cuddihy, R. Iwasaki, M. Mallison, C. Reed, J. Salit, A. Tracy, and **T. Marr**
Cold Spring Harbor Laboratory; Cold Spring Harbor, NY 11724
Marr: 516/367-8393, Fax: -8461, *marr@cshl.org* or *marr@cb.cshl.org*

Genome Topographer (GT) is an advanced genome informatics system that has received joint funding from DOE and NIH over a number of years. DOE funding has focused on GT tools supporting computational genome analysis, principally on sequence analysis. GT is scheduled for public release next spring under the auspices of the Cold Spring Harbor Human Genome Informatics Research Resource. GT has 17 major existing frameworks: 1. Views, including printing, 2. Default manager, 3. Graphical User Interface, 4. Query, 5. Project Manager, 6. Workspace Manager, 7. Asynchronous Process Manager, 8. Study Manager, 9. Help, 10. Application, 11. Notification, 12. Security, 13. World Wide Web Interface, 14. NCBI, 15. Reader, 16. Writer, 17. External Database Interface. GT Frameworks are independent sets of VisualWorks (client) or SmallTalkDB (GemStone) classes which interact to perform the duties required to satisfy the responsibilities of the specific framework. Each framework is clearly defined and has a well-defined interface to use it. These frameworks are used over and over in GT to perform similar duties in different places. GT has basic tools and special tools. Basic tools get used many times in different applications, while special tools tend to be special purpose, designed to do fairly limited things, although the distinction is somewhat arbitrary. Tools typically use several frameworks when they get assembled. Basic Tools: 1. Project Browser, 2. Editor/Viewer, 3. Query, 4. NCBI Entrez, 5. File reader/writer, 6. Map comparison, 7. Database Administrator, 8. Login, 9. Default, 10. Help. Special Tools: 1. Study Manager, 2. Compute Server, 3. Sequence Analysis, 4. Genetic Analysis. These frameworks and tools are combined with a comprehensive database schema of very rich biological expression linked with plugable computational tools. Taken together, these features allow users to construct, with relative ease, on-line databases of the primary data needed to study a genetic disease (or genes and phenotypes in general) from the stage of family collection and diagnostic ascertainment through cloning and functional analysis of candidate genes, including mutational analysis, expression information, and screening for biochemical interactions with candidate molecules. GT was designed on the premise that a highly informative, visual presentation of comprehensive data to a knowledgeable user is essential to their understanding. The advanced software engineering techniques that are promoted by using relatively new object oriented products has allowed GT to become a highly interactive and visually-oriented system that allows the user to concentrate on the problem rather than on the computer. Using the rich data representational features characteristic of this technology, the GT software enables users to construct models of real-world, complex biological phenomena. These unique features of GT are key to the thesis that such a system will allow users to discover otherwise intractable networks of interactions exhibited by complex genetic diseases.

The VisualWorks development environment allows the development of code that runs unchanged across all major workstation and personal computers, including PCS, Macintoshes and most Unix workstations.

DOE Grant No. DE-FG02-91ER61190.

A Flexible Sequence Reconstructor for Large-Scale DNA Sequencing: A Customizable Software System for Fragment Assembly

Gene Myers and Susan Larson
Department of Computer Science; University of Arizona; Tucson, AZ 85721
602/621-6612, Fax: -4246, *gene@cs.arizona.edu*
http://www.cs.arizona.edu/faktory

We have completed the design and begun construction of a software environment in support of DNA sequencing called the "FAKtory". The environment consists of (1) our previously described software library, FAK, for the core combinatorial problem of assembling fragments, (2) a Tcl/Tk based interface, and (3) a software suite supporting a modest database of fragments and a processing pipeline that includes clipping and vector prescreening modules. A key feature of our system is that it is highly customizable: the structure of the fragment database, the processing pipeline, and the operation of each phase of the pipeline are specifiable by the user. Such customization need only be established once at a given location, subsequently users see a relatively simple system tailored to their needs. Indeed one may direct the system to input a raw dataset of say ABI trace files, pass them through a customized pipeline, and view the resulting assembly with two button clicks.

The system is built on top of our FAK software library and as a consequence one receives (a) high-sensitivity overlap detection, (b) correct resolution to large high-fidelity repeats, (c) near perfect multi-alignments, and (d) support of constraints that must be satisfied by the resulting assemblies. The FAKtory assumes a processing pipeline for fragments that consists of an INPUT phase, any number and sequence of CLIP, PRESCREEN, and TAG phases, followed by an OVERLAP and then an ASSEMBLY phase. The sequence of clip, prescreen, and tag phases is customizable and every phase is controlled by a panel of user-settable preferences each of which permits setting the phase's mode to AUTO, SUPERVISED, or MANUAL. This setting determines the level of interaction required by the user when the phase is run, ranging from none to hands-on. Any diagnostic situations detected during pipeline processing are organized into a log that permits one to confirm, correct, or undo decisions that might have been made automatically.

The customized fragment database contains fields whose type may be chosen from TIME, TEXT, NUMBER, and WAVEFORM. One can associate default values for fields unspecified on input and specify a control vocabulary limiting the range of acceptable values for a given field (e.g., John, Joe, or Mary for the field Technician, and [1, 36] for the field Lane). This database may be queried with SQL-like predicates that further permit approximate matching over text fields. Common queries and/or sets of fragments selected by them may be named and referred to later by said name. The pipeline status of a fragment may be part of a query.

The system permits one to maintain a collection of alternative assemblies, to compare them to see how they are different, and directly manipulate assemblies in a fashion consistent with sequence overlaps. The system can be customized so that a priori constraints reflecting a given sequencing protocol (e.g. double-barreled or transposon-mapped) are automatically produced according to the syntax of the names of fragments (e.g. X.f and X.r for any X are mates for double-barreled sequencing). The system presents visualizations of the constraints applied to an assembly, and one may experiment with an assembly by adding and/or removing constraints. Finally, one may edit the multi-alignment of an assembly while consulting the raw waveforms. Special attention was given to optimizing the ergonomics of this time-intensive task.

DOE Grant No. DE-FG03-94ER61911.

The Role of Integrated Software and Databases in Genome Sequence Interpretation and Metabolic Reconstruction

Terry Gaasterland, Natalia Maltsev, **Ross Overbeek**, and Evgeni Selkov
Mathematics and Computer Science Division; Argonne National Laboratory; Argonne, IL 60439
630/252-4171, Fax: -5986, *gaasterl@mcs.anl.gov*
MAGPIE: *http://www.mcs.anl.gov/home/gaasterl/magpie.html*
WIT: *http://www.cme.msu.edu/WIT*

As scientists successfully sequence complete genomes, the issue of how to organize the large quantities of evolving sequence data becomes paramount. Through our work in comparative whole genome analysis (MAGPIE, Gaasterland) and metabolic reconstruction algorithms (WIT, Overbeek, Maltsev, and Selkov), we carry genome interpretation beyond the identification of gene products to customized views of an organism's functional properties.

MAGPIE is a system designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated.[1,2] DNA sequences produced in a sequencing project mature through a series of stages that each require different analysis activities. Even after DNA has been assembled into contiguous fragments and eventually into a single genome, it must be regularly reanalyzed. Any new data in public sequence databases may provide clues to the identity of genes. Over a year, for 2 megabases with 4-fold coverage, MAGPIE will request on the order of 100,000 outputs from remote analysis software, manipulate and manage the output, update the current analysis of the sequence data, and monitor the project sequence data for changes that initiate reanalysis.

In collaboration with Canada's Institute for Marine Biosciences and the Canadian Institute for Advanced Research, MAGPIE is being used to maintain and study comparative views of all open reading frames (ORFs) across fully sequenced genomes (currently 5), nearly completed genomes (currently 2) and 1 genome in progress (Sulfolobus solfataricus). Together, these genomes represent multiple archaeal and bacterial genomes and one eukaryotic genome. This analysis provides the necessary data to assign phylogenetic classifications to each ORF (e.g., "AE" for archaeal and eukaryotic). This data in turn provides the basis for validating and assessing functional annotations according to phylogenetic neighborhood (e.g., selecting the eukaryotic form of a biochemical function over a bacterial form for an "AE" ORF).[3]

Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact.To this end, the WIT (What Is There) system supports efforts to develop metabolic reconstructions. Such constructions, or models, are based on sequence data, clearly established biochemistry of specific organisms, understanding of the interdependencies of biochemical mechanisms. WIT thus offers a valuable tool for testing current hypotheses about microbial behavior. For example, a reconstruction may begin with a set of established enzymes (enzymes with strong similarities in identified coding regions to existing sequences for which the enzymatic function is known) and putative enzymes (enzymes with weak similarity to sequences of known function). From these initial "hits," within a phylogenetic perspective, we identify an initial set of pathways. This set can be used to generate a set of expected enzymes (enzymes that have not been clearly detected, but that would be expected given the set of hypothesized pathways) and missing enzymes (enzymes that occur in the pathways but for which no sequence has yet been biochemically identified for any organism). Further reasoning identifies tentative connective pathways.

In addition to helping curators develop metabolic reconstructions, WIT lets users examine models curated by experts, follow connections between more than two thousand metabolic diagrams, and compare models (e.g., which of certain genes that are conserved among bacterial genomes are found in higher life). The objective is to set the stage for meaningful simulations of microbial behavior and thus to advance our understanding of microbial biochemistry and genetics.

DOE Contract No. W-31-109-Eng-38 (ANL FWP No. 60427).

References

[1] T. Gaasterland and C. Sensen, Fully Automated Genome Analysis that Reflects User Needs and Preferences --a Detailed Introduction to the MAGPIE System Architecture, Biochemie, 78(4), (accepted).
[2] T. Gaasterland, J. Lobo, N. Maltsev, and G. Chen. Assigning Function to CDS Through Qualified Query Answering. In Proc. 2nd Int. Conf. Intell. Syst. for Mol. Bio., Stanford U. (1994).
[3] T. Gaasterland and E. Selkov. Automatic Reconstruction of Metabolic Structure from Incomplete Genome Sequence Data. In Proc. Int. Conf. Intell. Syst. for Mol. Bio., Cambridge, England (1995).

Database Transformations for Biological Applications

G. Christian Overton, Susan B. Davidson,[1] and **Peter Buneman**[1]
Department of Genetics and [1]Department of Computer and Information Science; University of Pennsylvania; Philadelphia, PA 19104
Overton: 215/573-3105, Fax: -3111, *coverton@cbil.humgen.upenn.edu*
Davidson: 215/898-3490, Fax: -0587, *susan@cis.upenn.edu*
Buneman: 215/898-7703, Fax: -0587,*peter@cis.upenn.edu*
http://agave.humgen.upenn.edu/cpl/cplhome.html
http://sdmc.iss.nus.sg/kleisli-stuff/MoreInfo.html

We have implemented a general-purpose query system, Kleisli, that provides access to a variety of "non-standard" data sources (e.g., ACeDB, ASN.1, BLAST), as well as to "standard" relational databases. The system represents a major advance in the ability to integrate the growing number and diversity of biology data sources conveniently and efficiently. It features a uniform query interface, the CPL query language, across heterogeneous data sources, a modular and extensible architecture, and most significantly for dealing with the Internet environment, a programmable optimizer. We have demonstrated the utility of the system in composing and executing queries that were considered difficult, if not unanswerable, without first either building a monolithic database or writing highly application-specific integration code (details and examples available at URL above).

In conjunction with other software developed in our group, we have assembled a toolset that supports a range of data integration strategies as well as the ability to create specialized data warehouses initialized from community databases. Our integration strategy is based upon the concept of "mediators", which serve a group of related applications by providing a uniform structural interface to the relevant data sources. This approach is cost-effective in terms of query development time and maintenance. We have examined in detail methods for optimizing queries such as "retrieve all known human sequence containing an Alu repeat in an intragenic region" where the data sources are heterogeneous and distributed across the Internet.

Transformation of data resources, that is the structural reorganization of a data resource from one form to another, arises frequently in genome informatics. Examples include the creation of data warehouses and database evolution. Implementing such transformations by hand on a case by case basis is time consuming and error prone. Consequently there is a need for a method of specifying, implementing and formally verifying transformations in a uniform way across a wide variety of different data models. Morphase is a prototype system for specifying transformations between data sources and targets in an intuitively appealing, declarative language based on Horn clause logic. Transformations specification in Morphase are translated into CPL and executed in the Kleisli system. The data-types underlying Morphase include arbitrarily nested records, sets, variants, lists and object identity, thus capturing the types common to most data formats relevant to genome informatics, including ASN.1 and ACE. Morphase can be connected to a wide variety of data sources simultaneously through Kleisli. In this way, data can be read from multiple heterogeneous data sources, transformed using Morphase according to the desired output format, and inserted into the target data source.

We have tested Morphase by applying it to a variety of different transformation problems involving Sybase, ACE and ASN.1. For example, we used it to specify a transformation between the Sanger Center's Chromosome 22 ACE database (ACE22DB) and a Chromosome 22 Sybase database (Chr22DB), as well as between a portion of GDB and Chr22DB. Some of these transformations had already been hand-coded without our tools, forming a basis for comparison.

Once the semantic correspondences between objects in the various databases were understood, writing the transformation program in Morphase was easy, even by a non-expert of the system. Furthermore, it was easy to find conceptual errors in the transformation specification. In contrast, the hand-coded programs were obtuse, difficult to understand, and even more difficult to debug.

DOE Grant No. DE-FG02-94ER61923.

Relevant Publications

P. Buneman, S.B. Davidson, K. Hart, C. Overton and L. Wong,"A Data Transformation System for Biological Data Sources," in Proceedings of VLDB, Sept. 1995 (Zurich, Switzerland). Also available as Technical Report MS-CIS-95-10, University of Pennsylvania, March 1995.

S.B. Davidson, C. Overton and P. Buneman, "Challenges in Integrating Biological Data Sources," J. Computational Biology 2 (1995), pp 557-572.

A. Kosky, "Transforming Databases with Recursive Data Structures," PhD Thesis, December 1995.

S.B. Davidson and A. Kosky, "Effecting Database Transformations Using Morphase," Technical Report MS-CIS-96-05, University of Pennsylvania.

A. Kosky, S.B. Davidson and P. Buneman, "Semantics of Database Transformations," Technical Report MS-CIS-95-25, University of Pennsylvania, 1995.

K. Hart and L. Wong, "Pruning Nested Data Values Using Branch Expressions With Wildcards," In Abstracts of MIMBD, Cambridge, England, July 1995.

Las Vegas Algorithm for Gene Recognition: Suboptimal and Error-Tolerant Spliced Alignment

Sing Hoi Sze and **Pavel A. Pevzner**[1]
Departments of Computer Science and [1]Mathematics; University of Southern California; Los Angeles, CA 90089
Pevzner: 213/740-2407, Fax: -2424
ppevzner@hto.usc.edu
http://www-hto.usc.edu/software/procrustes

Recently, Gelfand, Mironov, and Pevzner (Proc. Natl. Acad. Sci. USA, 1996, 9061-9066) proposed a spliced alignment approach to gene recognition that provides 99% accurate recognition of human gene if a related mammalian protein is available. However, even 99% accurate gene predictions are insufficient for automated sequence annotation in large-scale sequencing projects and therefore have to be complemented by experimental gene verification. 100% accurate gene predictions would lead to a substantial reduction of experimental work on gene identification. Our goal is to develop an algorithm that either predicts an exon assembly with accuracy sufficient for sequence annotation or warns a biologist that the accuracy of a prediction is insufficient and further experimental work is required. We study suboptimal and error-tolerant spliced alignment problems as the first steps towards such an algorithm, and report an algorithm which provides 100% accurate recognition of human genes in 37% of cases (if a related mammalian protein is available). For 52% of genes, the algorithm predicts at least one exon with 100% accuracy.

DOE Grant No. DE-FG03-97ER62383.

Foundations for a Syntactic Pattern-Recognition System for Genomic DNA Sequences: Languages, Automata, Interfaces, and Macromolecules

David B. Searls and G. Christian Overton[1]
SmithKline Beecham Pharmaceuticals; King of Prussia, PA 19406
610/270-4551, Fax: -5580, *searldb@sb.com*
[1]Department of Genetics; University of Pennsylvania; Philadelphia, PA 19104

Viewed as strings of symbols, biological macromolecules can be modelled as elements of formal languages. Generative grammars have been useful in molecular biology for purposes of syntactic pattern recognition, for example in the author's work on the GenLang pattern matching system, which is able to describe and detect patterns that are probably beyond the capability of a regular expression specification. More recently, grammars have been used to capture intramolecular interactions or long-distance dependencies between residues, such as those arising in folded structures. In the work of Haussler and colleagues, for example, stochastic context-free grammars have been used as a framework for "learning" folded RNA structures such as tRNAs, capturing both primary sequence information and secondary structural covariation. Such advances make the study of the formal status of the language of biological macromolecules highly relevant, and in particular the finding that DNA is beyond context-free has already created challenges in algorithm design.

Moreover, to date, such methods have not been able to capture relationships between strings in a collection, such as those that arise via intermolecular interactions, or evolutionary relationships implicit in alignments. Recently we have attempted to remedy this by showing (1) how formal grammars can be extended to describe interacting collections of molecules, such as hybridization products and, potentially, multimeric or physiological protein interactions, and (2) how simple automata can be used to model evolutionary relationships in such a way that complex model-based alignment algorithms can be automatically generated by means of visual programming. These results allow for a useful generalization of the language-theoretic methods now applied to single molecules.

In addition, we describe a new software package—bioWidget—for the rapid development and deployment of graphical user interfaces (GUIs) designed for the scientific visualization of molecular, cellular and genomics information. The overarching philosophy behind bioWidgets is componentry: that is, the creation of adaptable, reusable software, deployed in modules that are easily incorporated in a variety of applications and in such a way as to promote interaction between those applications. This is in sharp distinction to the common practice of developing dedicated applications. The bioWidgets project additionally focuses on the development of specific applications based on bioWidget componentry, including chromosomes, maps, and nucleic acid and peptide sequences.

The current set of bioWidgets has been implemented in Java with the goal in mind of delivering local applications and distributed applets via Intranet/Internet environments as required. The immediate focus is on developing interfaces for information stored in distributed heterogeneous databases such as GDB, GSDB, Entry, and ACeDB. The issues we are addressing are database access, reflecting database schemas in bioWidgets, and performance. We are also directing our efforts into creating a consortium of bioWidget developers and end-users. This organization will create standards for and encourage the development of bioWidget components. Primary participants in the consortium include Gerry Rubin (UC Berkeley) and Nat Goodman (Jackson Labs).

DOE Grant No. DE-FG02-92ER61371.

Relevant Publications

D.B. Searls, "String Variable Grammar: A Logic Grammar Formalism for DNA Sequences," Journal of Logic Programming 24 (1,2):73-102 (1995).

D.B. Searls, "Formal Grammars for Intermolecular Structure," First International Symposium on Intelligence in Neural and Biological Systems, 30-37 (1995).

D.B. Searls and K.P. Murphy, "Automata-Theoretic Models of Mutation and Alignment," Third International Conference on Intelligent Systems for Molecular Biology, 341-349 (1995).

D.B. Searls, "bioTk: Componentry for Genome Informatics Graphical User Interfaces," Gene 163 (2):GC1-16 (1995).

Analysis and Annotation of Nucleic Acid Sequence

David J. States, Ron Cytron, Pankaj Agarwal, and Hugh Chou
Institute for Biomedical Computing; Washington University; St. Louis, MO 63108
314/362-2134, Fax: -0234, *states@ibc.wustl.edu*
http://www.ibc.wustl.edu

Bayesian estimates for sequence similarity: There is an inherent relationship between the process of pairwise sequence alignment and the estimation of evolutionary distance. This relationship is explored and made explicit. Assuming an evolutionary model and given a specific pattern of observed base mismatches, the relative probabilities of evolution at each evolutionary distance are computed using a Bayesian framework. The mean or the median of this probability distribution provides a robust estimate of the central value. Bayesian estimates of the evolutionary distance incorporate arbitrary prior information about variable mutation rates both over time and along sequence position, thus requiring only a weak form of the molecular-clock hypothesis.

The endpoints of the similarity between genomic DNA sequences are often ambiguous. The probability of evolution at each evolutionary distance can be estimated over the entire set of alignments by choosing the best alignment at each distance and the corresponding probability of duplication at that evolutionary distance. A central value of this distribution provides a robust evolutionary distance estimate. We provide an efficient algorithm for computing the parametric alignment, considering evolutionary distance as the only parameter.

These techniques and estimates are used to infer the duplication history of the genomic sequence in C. elegans and in S. cerevisae. Our results indicate that repeats discovered using a single scoring matrix show a considerable bias in subsequent evolutionary distance estimates.

Model based sequence scoring metrics: PAM based DNA comparison metric has been extended to incorporate biases in nucleotide composition and mutation rates, extending earlier work (States, Gish and Altschul, 1993). A codon based scoring system has been developed that incorporates the effects biased codon utilization frequencies.

A dynamic programming algorithm has been developed that will optimally align sequences using a choice of comparison measures (non-coding vs. coding, etc.). We are in the process of evaluating this approach as a means for identifying likely coding regions in cDNA sequences.

Efficient sequence similarity search tools: Most sequence search tools have been designed for use with protein sequence queries a few hundred residues long. The analysis of genomic DNA sequence necessitates the use of queries hundreds of kilobases or even megabases in length. A memory and computationally efficient search tool has been developed for the identification of repeats and sequence similarity in very large segments of nucleic acid sequence. The tool implements optimal encoding of the word table, repeat filters, flexible scoring systems, and analytically parameterized search sensitivity. Output formats are designed for the presentation of genomic sequence searches.

Federated databases: A sybase server and mirror for GSDB are being developed to facilitate the annotation of repeat sequence elements in public data repositories.

DOE Grant No. DE-FG02-94ER61910.

Gene Recognition, Modeling, and Homology Search in GRAIL and genQuest

Ying Xu, Manesh Shah, J. Ralph Einstein, Sherri Matis, Xiaojun Guan, Sergey Petrov, Loren Hauser,[1] Richard J. Mural,[1] and **Edward C. Uberbacher**
Computer Science and Mathematics and [1]Biology Divisions; Oak Ridge National Laboratory; Oak Ridge, TN 37831
Uberbacher: 423/574-6134, Fax: -7860, *ube@ornl.gov*
http://compbio.ornl.gov

GRAIL is a modular expert system for the analysis and characterization of DNA sequences which facilitates the recognition of gene features and gene modeling. A new version of the system has been created with greater sensitivity for exon prediction (especially in AT rich regions), more accurate splice site prediction, and robust indel error detection capability. GRAIL 1.3 is available to the user in a Motif graphical client-server system (XGRAIL), through WWW-Netscape, by e-mail server, or callable from other analysis programs using Unix sockets.

In addition to the positions of protein coding regions and gene models, the user can view the positions of a number of other features including poly-A addition sites, potential Pol II promoters, CpG islands and both complex and simple repetitive DNA elements using algorithms developed at ORNL. XGRAIL also has a direct link to the genQuest server, allowing characterization of newly obtained sequences by homology-based methods using a number of protein, DNA, and motif databases and comparison methods such as FastA, BLAST, parallel Smith-Waterman, and special algorithms which consider potential frameshifts during sequence comparison.

Following an analysis session, the user can use an annotation tool which is part of the XGRAIL 1.3 system to generate a “feature table” report describing the current sequence and its properties. Links to the GSDB sequence database have been established to record computer-based analysis of sequences during submission to the database or as third party annotation.

Gene Modeling and Client-Server GRAIL: In addition to the current coding region recognition capabilities based on a multiple sensor-neural network and rule base, modules for the recognition of features such as splice junctions, transcription and translation start and stop, and other control regions have been constructed and incorporated into an expert system (GAP III) for reliable computer-based modeling of genes. Heuristic methods and dynamic programming are used to construct first pass gene models which include the potential for modification of initially predicted exons. These actions result in a net improvement in gene characterization, particularly in the recognition of very short coding regions. Translation of gene models and database searches are also supported through access to the genQuest server (described below).

Model Organism Systems: A number of model organism systems have been designed and implemented and can be accessed within the XGRAIL 1.3 client including Escherichia coli, Drosophila melanogaster and Arabidopsis thaliana. The performance of these systems is basically equivalent to the Human GRAIL 1.3 system. Additional model organism systems, including several important microorganisms, are in progress.

Error Detection in Coding Sequences: Single-pass DNA sequencing is becoming a widely used technique for gene identification from both cDNA and genomic DNA sequences. An appreciably higher rate of base insertion and deletion errors (indels) in this type of sequence can cause serious problems in the recognition of coding regions, homology search, and other aspects of sequence interpretation. We have developed two error detection and "correction" strategies and systems which make low-redundancy sequence data more informative for gene identification and characterization purposes. The first algorithm detects sequencing errors by finding changes in the statistically preferred reading frame within a possible coding region and then rectifies the frame at the transition point to make the potential exon candidate frame-consistent. We have incorporated this system in GRAIL 1.3 to provide analysis which is very error tolerant. Currently the system can detect about 70% of the indels with an indel rate of 1%, and GRAIL identifies 89% of the coding nucleotides compared to 69% for the system without error correction. The algorithm uses dynamic programming and runs in time and space linear to the size of the input sequence.

In the second method, a Smith-Waterman type comparison is facilitated in which the frame of DNA translation to protein sequence can change within the sequence. The transition points in the translation frame are determined during the comparison process and a best match to potential protein homologs is obtained with sections of translations from more than one frame. The algorithm can detect homologies with a sensitivity equivalent to Smith-Waterman in the presence of 5% indel errors.

Detection of Regulatory Regions: An initial Polymerase II promoter detection system has been implemented which combines individual detectors for TATA, CAAT, GC, cap, and translation start elements and distance information using a neural network. This system finds about 67% of TATA containing promoters with a false positive rate of one per 35 kilobases. Additionally a systems to detect potential polyA addition sites and CpG islands has been incorporated into GRAIL.

The GenQuest Sequence Comparison Server: The genQuest server is an integrated sequence comparison server which can be accessed via e-mail, using Unix sockets from other applications, Netscape, and through a Motif graphical client-server system. The basic purpose of the server system is to facilitate rapid and sensitive comparison of DNA and protein sequences to existing DNA, protein, and motif databases. Databases accessed by this system include the daily updated GSDB DNA sequence database, SwissProt, the dbEST expressed sequence tag database, protein motif libraries and motif analysis systems (Prosite, BLOCKS), a repetitive DNA library (from J. Jurka), Genpept, and sequences in the PDB protein structural database. These options can also be accessed from the XGRAIL graphical client tool.

The genQuest server supports a variety of sequence query types. For searching protein databases, queries may be sent as amino acid or DNA sequence. DNA sequence can be translated in a user specified frame or in all 6 frames. DNA-DNA searches are also supported. User selectable methods for comparison include the Smith-Waterman dynamic programming algorithm, FastA, versions of BLAST, and the IBM dFLASH protein sequence comparison algorithm. A variety of options for search can be specified including gap penalties and option switches for Smith-Waterman, FastA, and BLAST, the number of alignments and scores to be reported, desired target databases for query, choice of PAM and Blosum matrices, and an option for masking out repetitive elements. Multiple target databases can be accessed within a single query.

Additional Interfaces and Access: Batch GRAIL 1.3 is a new "batch" GRAIL client allows users to analyze groups of short (300-400 bp) sequences for coding character and automates a wide choice of database searches for homology and motifs. A Command Line Sockets Client has been constructed which allows remote programs to call all the basic analysis services provided by the GRAIL-genQuest system without the need to use the XGRAIL interface. This allows convenient integration of selected GRAIL analyses into automated analysis pipelines being constructed at some genome centers. An XGRAIL Motif Graphical Client for the GRAIL release 1.3 has been constructed using Motif with versions for a wide variety of UNIX platforms including Sun, Dec, and SGI. The e-mail version of GRAIL can be accessed at grail@ornl.gov and the e-mail version of genQuest can be accessed at Q@ornl.gov. Instructions can be obtained by sending the word "help" to either address. The Motif or Sun versions of XGRAIL, batch GRAIL, and XgenQuest client software are available by anonymous ftp from grailsrv.lsd.ornl.gov (124.167.140.21). Both GRAIL and genQuest are accessible over the World Wide Web (URL http://compbio.ornl.gov). Communications with the GRAIL staff should be addressed to GRAILMAIL@ornl.gov.

DOE Contract No. DE-AC05-840R21400.

Informatics Support for Mapping in Mouse-Human Homology Regions

Edward Uberbacher, Richard Mural,[1] Manesh Shah, Loren Hauser,[1] and Sergey Petrov
Computer Science and Mathematics Division and [1]Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831
423/574-6134, Fax: -7860, *ube@ornl.gov*

The purpose of this project is to develop databases and tools for the Oak Ridge National Laboratory (ORNL) Mouse-Human Mapping Project, including the construction of a mapping database for the project; tools for managing and archiving cDNAs and other probes used in the laboratory; and analysis tools for mapping, interspecific backcross, and other needs. Our initial effort involved installing and developing a relational SYBASE database for tracking samples and probes, experimental results, and analyses. Recent work has focused on a corresponding ACeDB implementation containing mouse mapping data and providing numerous graphical views of this data. The initial relational database was constructed with SYBASE using a schema modeled on one implemented at the Lawrence Livermore National Laboratory (LLNL) center; this was because of documentation available for the LLNL system and the opportunity to maximize compatibility with human chromosome 19 mapping. (Major homologies exist between human chromosome 19 and mouse chromosome 7, the initial focus of the ORNL work.)

With some modification, our ACeDB implementation was modeled somewhat on the Lawrence Berkeley National Laboratory (LBNL) chromosome 21 ACeDB system and designed to contain genetic and physical mouse map data as well as homologous human chromosome data. The usefulness of exchanging map information with LLNL (human chromosome 19) and potentially with other centers has led to the implementation of procedures for data export and the import of human mapping data into ORNL databases.

User access to the system is being provided by workstation forms-based data entry and ACeDB graphical data browsing. We have also implemented the LLNL database browser to view human chromosome 19 data maintained at LLNL, and arrangements are being made to incorporate mouse mapping information into the browser. Other applications such as the *Encyclopedia of the Mouse*, specific tools for archiving and tracking cDNAs and other mapping probes, and analysis of interspecific backcross data and YAC restriction mapping have been implemented.

We would like to acknowledge use of ideas from the LLNL and LBNL Human Genome Centers.

DOE Contract No. DE-AC05-840R21400.

SubmitData: Data Submission to Public Genomic Databases

Manfred D. Zorn
Software Technologies and Applications Group; Information and Computing Sciences Division; Lawrence Berkeley National Laboratory; University of California; Berkeley CA 94720
510/486-5041, Fax: -4004, *mdzorn@lbl.gov*
http://www-hgc.lbl.gov/submitr.html

Making information generated by the various genome projects available to the community is very important for the researcher submitting data and for the overall project to justify the expenses and resources. Public genome databases generally provide a protocol that defines the required data formats and details how they accept data, e.g., sequences, mapping information. These protocols have to strike a balance between ease of use for the user and operational considerations of the database provider, but are in most cases rather complex and subject to change to accommodate modifications in the database.

SubmitData is a user interface that formats data for submission to GSDB or GDB. The user interface serves data entry purposes, checking each field for data types, allowed ranges and controlled values, and gives the user feedback on any problems. Besides one-time submissions, templates can be created that can later be merged with TAB-delimited data files, e.g., as produced by common spreadsheet programs. Variables in the template are then replaced by values in defined columns of the input data file. Thus submitting large amounts of related data becomes as easy as selecting a format and supplying an input filename. This allows easy integration of data submission into the data generation process.

The interface is generated directly from the protocol specifications. A specific parser/compiler interprets the protocol definitions and creates internal objects that form the basis of the user interface. Thus a working user interface, i.e., static layout of buttons and fields, data validation, is automatically generated from the protocol definitions. Protocol modifications are propagated by simply regenerating the interface.

The program has been developed using ParcPlace VisualWorks and currently supports GSDB, GDB and RHdb data submissions. The program has been updated to use VisualWorks 2.0.

DOE Contract No. DE-AC03-76SF00098.

The Human Genome: Science and the Social Consequences; Interactive Exhibits and Programs on Genetics and the Human Genome

Charles C. Carlson
The Exploratorium; San Francisco, CA 94123
415/561-0319, Fax: -0307; *charliec@exploratorium.edu*

From April through September 1995, the Exploratorium mounted a special exhibition called *Diving into the Gene Pool* consisting of 26 interactive exhibits developed over the course of three years. The exhibits introduce the science of genetics and increase public awareness of the Human Genome Project and its implications for society. Founded in the success of exhibits developed for the 1992 genetics and biotechnology symposium "Winding Your Way Through DNA" (co-hosted with the University of California, San Francisco), the 1995 exhibition aimed to create an engaging and accessible presentation of specific information about genetic science and our understanding of the structure and function of the human genome, genetic technology, and ethical issues surrounding current genetic science.

In addition to creating a unique collection of exhibits, the project developed a range of supplemental public programming to provide public forum for discussion and interaction about genetics and bioethics. A lecture series entitled "Bioethics and the Human Genome Project," featured such key thinkers as Mary Claire King, Leroy Hood, David Martin, Troy Duster, Michael Yesley, William Atchley, and Joan Hamilton (among others). A weekend event program focused on biodiversity in animal and plant life with events such as "Seedy Science," "Blooming Genes," and "Dog Diversity." A Biotech Weekend offered access to new technologies through demonstrations by local biotech firms and genetic counselors. And a specially-commissioned theatre piece, "Dog Tails," provided a instructive and comic look for kids into the foundations of genetics and issues of diversity.

In the 5-month exhibition period, approximately 300,000 visitors had the opportunity to visit the exhibition, and well over 5,000 participated in the special programming. Following the exhibition's close, the new exhibits will become a permanent part of the Exploratorium's collection of over 650 interactive exhibits.

Additional funding for 1995-96 will support formal outside evaluation of the effectiveness of the exhibits, and support exhibit remediation based on the evaluation findings. This activity will both strengthen the Exploratorium's permanent collection of genetics exhibits and help to develop a feasibility study for a travelling version of the genetics exhibition for other museums around the country and the world.

DOE Grant No. DE-FG03-93ER61583.

Documentary Series for Public Broadcasting

Graham Chedd and **Noel Schwerin**
Chedd-Angier Production Company; Watertown, MA 02172
617/926-8300, Fax: -2710

Designed as a 4-hour documentary series for Public Broadcasting, *Genetics in Society* (working title) will explore the ethical, legal, and social implications of genetic technology. Currently funded and in production for a 90-minute special (*Testing Family Ties*), the first program profiles several individuals and families as they confront genetic tests and the information they generate. One high-risk cancer family struggles to make sense of their genetic legacy as it debates prophylactic surgery and whether or not to test for *BRCA*1 and *BRCA*2. In a family without that family risk, news of the Ashkenazi *BRCA*1 finding pushes an anxious Jewish woman to demand testing for herself and her young daughter. In another, a woman chooses to carry to term her prenatally diagnosed Cystic Fibrosis twins, despite social and personal pressures. In a third, a scientist researching the so-called "obesity gene" at a biotech company debates the proper "marketing" of his research and confronts the larger questions it raises about what should be considered "normal" and what constitutes therapy vs enhancement.

Testing Family Ties will explore not only what genetic technology does—in testing, drug development, and potential therapy—but what it means to our sense of self, family, and future and to our concepts of health and normality.

Depending on outstanding funding requests, *Genetics in Society* will be broadcast in the Fall of 1996 or the Winter of 1997 on PBS. Noel Schwerin is Producer/Director. Graham Chedd is Executive Producer.

DOE Grant No. DE-FG06-95ER61995.

Human Genome Teacher Networking Project

Debra L. Collins and **R. Neil Schimke**
Genetics Education Center; Division of Endocrinology and Genetics; University of Kansas Medical Center; Kansas City, KS 66160-7318
913/588-6043, Fax: -4060, *collins@ukanvm.cc.ukans.edu*
http://www.kumc.edu/GEC

This project links over 150 middle and secondary teachers from throughout the United States with genetic and public policy professionals, as well as families who are knowledgeable about the ethical, legal, and social implications

(ELSI) of the Human Genome Project. Teachers network with peers and professionals, and acquire new sources of information during four phases: 1) the first one-week summer workshop to update teachers on human genetics concepts and new sources for classroom curricula including online resources; 2) classroom use of new materials and information; 3) the second one-week summer workshop where teachers return to exchange successful teaching ideas and plan peer teaching sessions and mentor networking; 4) dissemination of genetic information through in-services and workshops for colleagues; and collaboration with genetic professional participating in our Mentor Network.

The applications of Human Genome Project technology are emphasized. Individuals who have contact and experience with patients, including clinical geneticists, genetic counselors, attorneys, laboratories geneticists and families, take part in didactic sessions with teachers. Throughout the workshop, family panels provide an opportunity for participants to compare their textbook-based knowledge of genetic conditions with the personal experiences of families who discuss their condition, including: diagnosis, treatment, genetic risk, decisions, insurance, employment, family planning, and confidentiality.

Because of this project, teachers feel more prepared and confident teaching about human genetics, the Human Genome Project, and ELSI topics. The teachers are effective in disseminating knowledge of genetics to their students who show a significant increase in human genome knowledge compared to students whose teachers have not participated in this project.

Teacher dissemination activities extend the project beyond participation at summer workshops. To date, 55 workshop participants have completed all four project phases by organizing more than 200 local, regional, and national teacher education programs to disseminate knowledge and resources. More than 1500 colleagues and the general public have participated in teacher workshops, and over 56,000 students have been reached through project participants and their peers.

The project participants organize interdisciplinary peer teaching sessions including bioethical decision making sessions combining debate and biology classes; sessions for social studies teachers; human genetics and multi-cultural collaborations; cooperative learning activities; and curricular development sessions. Students were involved in sessions on ethics, politics, economics and law. Teachers organize bioethics curriculum writing sessions, laboratory activities using electrophoresis as well as other biotechnology, and sessions on genetic databases.

A World Wide Web home page for Genetics Education assists teachers in remaining current on genetic information and helps them find answers to student inquiries. The home page has links to numerous genome sites, sources of information on genetic conditions, networking opportunities with other genetics education programs, teaching resources, lesson plan ideas, and the Mentor Network of genetic professionals and a network of family support groups willing to work with teachers and their students.

DOE Grant No. DE-FG02-92ER61392.

Human Genome Education Program

Lane Conn
Human Genome Education Program; Stanford Human Genome Center; Palo Alto, CA 94304
415/812-2003, Fax: -1916, *lconn@toolik.stanford.edu*

The Human Genome Education Program (HGEP) operates within the Stanford Human Genome Center. It is a collaborative effort among HGEP staff, Genome Center scientists, collaborating staff from other education programs, experienced high school teachers, and an Advisory Panel in the fields of science, education, social science, assessment, and ethics.

The Human Genome Project will have a profound impact on society with its applications in testing for and improving treatment of genetic disease and the many uses of DNA profiling. The goal of HGEP is to help prepare high school students and community members to be able to make educated decisions on the personal, ethical, social and policy questions raised by the application of genome information and technology in their lives.

The primary objectives for HGEP are to (1) develop a human genome curriculum for high school science and (2) education outreach to schools and community groups in the San Francisco Bay Area. To achieve Objective 1, the HGEP is working to develop, field test, and prepare for national dissemination a two laboratory-based curriculum units for high school students. Unit 1, “Dealing With Genetic Disorders,” explores the variety of treatment options potentially available for a genetic disorder, including gene therapy. Unit 2, “DNA Snapshots, Peeking at Your DNA,” explores human relatedness through examining the student’s own DNA polymorphisms using PCR.

Each unit is centered around a societal or ethical problem raised by these important applications of genome information and technology. Students use modeling exercises and inquiry laboratory experiments to learn about the science behind a given application. Students then combine the science they have learned with other relevant information to choose a solution to the societal/ethical problem posed in the unit. As a culminating activity, the students work in groups to present and defend their solution.

To achieve Objective 2, the HGEP provides Genome Center tours for teacher, student and community groups that involve pre-tour lectures; tour exploration of genome mapping, sequencing and informatics; and post-tour lecture and discussion on genome applications, and their social and ethical implications. Also, the education program continues to work to establish and sustain local science education partnerships among schools, industry, universities and national laboratories.

DOE Grant No. DE-FG03-96ER62161.

Your World/Our World–Biotechnology & You: Special Issue on the Human Genome Project

Jeff Davidson and **Laurence Weinberger**
Pennsylvania Biotechnology Association; State College, PA 16801
814/238-4080, Fax: -4081, *73150.1623@compuserve.com*

Your World/Our World is a biotechnology science magazine published semi-annually by the non-profit Pennsylvania Biotechnology Association (PBA) describing for seventh to tenth grade students the excitement and achievements of contemporary biotechnology. This is the only continuing source of biotechnology education specifically directed to this age group - an age at which students too frequently are turned off from science. The special Spring 1996 issue will be devoted to the presentation of the science behind the HGP, the HGP itself, and the ethical, legal, and social issues generated by the project. The strong emphasis on attractive graphic presentation and age appropriate text that have been the hallmark of the earlier issues, which have been highly acclaimed and well received by the educational, scientific, and business community, will be continued.

PBA believes that increased educational opportunities to learn about biotechnology are most effective if presented at the seventh to tenth grade levels for the following reasons:

- Full semester life science and biology classes often occur for the first time in these grades;
- Across the nation, textbooks are typically 10 to 14 years old, and even the most recent textbooks are quickly dated by the rapid development in the biological sciences;
- Curricula at this level are more flexible than high school curricula, allowing the addition of information about exciting biological developments; and
- Science at this level is generally not elective, and, therefore, a very comprehensive student population is addressed rather than the more selective populations available later in the educational program.

In creating *Your World/Our World*, the PBA defined the following educational goals to guide the development of the magazine:

- Contribute to general science literacy and an educated electorate;
- Contribute to biological and technological literacy; and
- Motivate students to pursue additional science study and careers in science, particularly among women and minority populations.

PBA recognizes that it has been a point of pride that biotechnologists have been uniquely concerned with the impact of their technology on society and have been the first to raise and encourage responsible public debate without being forced to do so by others. To do less now for the children would be a breach of this responsible history. Accordingly, this special HGP issue will address the ethical, legal, and social issues raised by the new genomic technologies. Special ethics advisors have been recruited to aid in the development of these aspects.

A complimentary copy of the special issue and its teachers' guide will be mailed to every public and private school seventh to tenth grade science teacher (approximately 40,000) in the United States. A cover announcement will explain the origin and development of the magazine and of the special edition. Teachers will be invited to purchase full classroom packets (30 copies & teacher's guide) from the PBA, but, if they are not able to afford the packets, they will be asked to respond by postcard indicating their interest. The cost of the packets will probably be in the $20 range. The PBA is actively seeking additional support so that the issue may be distributed for free or at a reduced cost. In addition, parts of the special issue will be available over the Internet via a World Wide Web Page.

PBA believes this is a unique opportunity to educate America's youth about the HGP and insure that accurate non-sensational information will be made available to our country's children.

DOE Grant No. DE-FG02-95ER62107.

The Human Genome Project and Mental Retardation: An Educational Program

Sharon Davis
Department of Research and Program Services; The Arc of the United States; Arlington, TX 76010
817/261-6003, Fax: /277-3491, *sdavis@metronet.com*
http://The Arc.org/welcome.html

The Arc of the United States, a national organization on mental retardation, with 140,000 members and more than 1000 affiliated chapters proposes to educate its general membership and volunteer leaders about the Human Genome Project as it relates to mental retardation. A large number of identified causes of mental retardation are genetic, and many family members of The Arc deal with issues related to a genetic condition on a daily basis. We believe it is critical for our members and leaders to be educated about the scientific and ethical, legal and social aspects of the HGP, so that the association can evaluate and discuss the issues and develop positions based on adequate knowledge.

The major objectives of the proposed three-year project are to develop and disseminate educational materials for members/leaders of The Arc to inform them about the Human Genome Project and mental retardation and to conduct training on the scientific and ethical, legal and social aspects of the Human Genome Project and mental retardation using The Arc's existing training vehicles.

The Arc will develop and disseminate educational materials oriented toward families and conduct training at its national and state conventions, local chapter meetings and at board of director's meetings. The American Association of University Affiliated Programs for Persons with Developmental Disabilities (AAUAP) will assist with the project by providing needed expertise. The AAUAP membership includes university faculty who are experts on the genetic causes of mental retardation and on related ethical, legal and social issues. An advisory panel of university scientists and leaders of The Arc will guide the project.

DOE Grant No. DE-FG03-96ER62162.

Pathways to Genetic Screening: Molecular Genetics Meets the High-Risk Family

Troy Duster and Diane Beeson[1]
Institute for the Study of Social Change; University of California; Berkeley, CA 94705
510/642-0813, Fax: /8674, *nitrogn@violet.berkeley.edu*
[1]Department of Sociology; California State University; Hayward, CA 94542

The proliferation of genetic screening and testing is requiring increasing numbers of Americans to integrate genetic knowledge and interventions into their family life and personal experience. This study examines the social processes that occur as families at risk for two of the most common autosomal recessive diseases, sickle cell disease (SC) and cystic fibrosis (CF), encounter genetic testing. Since each of these diseases is found primarily in a different ethnic/racial group (CF in European Americans and SC is African Americans), this research will clarify the role of culture in integrating genetic testing into family life and reproductive planning. A third type of genetic disorder, the thalassemias, has recently been added to our sample in order to extend our comparative frame to include other ethnic and racial groups. In California, the thalassemias primarily affect Southeast Asian immigrants, although another risk group is from the Mediterranean region. Thalassemias, like cystic fibrosis and sickle cell disease, have a similar pattern of inheritance and raise similarly serious bio-medical challenges and issues of information management.

Data are drawn from interviews with members of families in which a gene for CF, SC or thalassemia has been identified. Data collection consists primarily of focused interviews with approximately 400 individuals from families in which at least one member has been identified as having a genetic disorder (or trait). In the most recent phase of the research, we are conducting focus groups selected to achieve stratified homogeneity around key social dimensions such as gender and relationship to disease. This is clarifying the social processes that facilitate and inhibit genetic testing.

We are currently assessing the concerns expressed by respondents about the potential uses of genetic information. We find strong patterns of concern, often based on personal experience, that genetic information may be used in ways that family members perceive as dangerous and/or discriminatory. First among these concerns is fear of losing access to health care. Additional concerns include fear of genetic discrimination in employment and other types of insurance, particularly life insurance. Similar patterns of concern exist among members of each ethnic group, and are frequently the focus of attention among family members, but take somewhat different form within each cultural group. These concerns constitute a growing obstacle to widespread use of genetic testing.

DOE Grant No. DE-FG03-92ER61393.

Intellectual Property Issues in Genomics

Rebecca S. Eisenberg
University of Michigan Law School; Ann Arbor, MI 48109
313/763-1372, Fax: -9375, *rse@umich.edu*

Intellectual property issues have been uncommonly salient in the recent history of advances in genomics. Beginning with the filing of patent applications by NIH on the first batch of expressed sequence tags (ESTs) from the laboratory of Dr. Craig Venter, each new development has been met with speculation about its strategic significance from an intellectual property perspective. Are ESTs of unknown function patentable, or is further work necessary before they satisfy patent law standards? Will patents on such fragments promote commercial investment in product development, or will they interfere with scientific communication and collaboration and retard the overall research effort? Without patent rights, how may the owners of private cDNA sequence databases earn a return on their investment while still permitting other investigators to obtain access to the information on reasonable terms? What are the rights of those who contribute resources such as cDNA libraries that are used to create the databases, and of those who identify sequences of interest out of the morass of information in the databases by formulating appropriate queries? Will the disclosure of ESTs in the public domain preclude patenting of subsequently characterized full-length genes and gene products? And why would a commercial firm invest its own resources in generating an EST database for the public domain?

Two factors have contributed to the fascination with intellectual property in this setting. First is a perception that some pioneers in genomics have sought to claim intellectual property rights that reach beyond their actual achievements to cover future discoveries yet to be made by others. For example, the controversial NIH patent applications claimed rights not only in the ESTs that were actually set forth in the specifications, but also in the full-length cDNAs that might be obtained by using the ESTs as probes, as well as in other, undisclosed fragments of those genes. More recently, private owners of cDNA sequence databases have set as a condition for access agreement to offer the database owners licenses to any resulting intellectual property. These efforts to claim rights to the future discoveries of others raise issues about the fairness and efficiency of the law in allocating rewards and incentives along the path of cumulative innovation.

Second is the counterintuitive alignment of interests in the debate. It was a public institution, NIH, that initially favored patenting discoveries that some representatives of industry thought should remain unpatented, and it was a major pharmaceutical firm, Merck & Co., that ultimately took upon itself the quasi-governmental function of sponsoring a university-based effort to place comparable information in the public domain. These topsy-turvy positions in the public and private sectors raise intriguing questions about the proper roles of government and industry in genomics research, and about who stands to benefit (and who stands to lose) from the private appropriation of genomic information.

DOE Grant No. DE-FG02-94ER61792.

AAAS Congressional Fellowship Program

Stephen Goodman
The American Society of Human Genetics; Bethesda, MD 20814-3998
301/571-1825, Fax: /530-7079, *society@genetics.faseb.org*

Few individuals in the genetics community are conversant with federal mechanisms for developing and implementing policy on human genetics research. In 1995 the American Society of Human Genetics (ASHG), in conjunction with DOE, initiated an American Association for the Advancement of Science (AAAS) Congressional Fellowship Program to strengthen the dialogue between the professional genetics community and federal policymakers. The fellowship will allow genetics professionals to spend a year as special legislative assistants on the staff of members of Congress or on congressional committees. Directed toward productive scientists, the program is intended to attract independent investigators.

In addition to educating the scientific community about the public policy process, the fellowship is expected to demonstrate the value of science-government interactions and make practical contributions to the effective use of scientific and technical knowledge in government. The program includes an orientation to legislative and executive operations and a year-long weekly seminar on issues involving science and public policy.

Unlike similar government programs, this fellowship is aimed primarily at scientists outside government. It emphasizes policy-oriented public service rather than observational learning and designates its fellows as free agents rather than representatives of their sponsoring societies.

One of the goals of DOE and ASHG is to develop a group of nongovernmental professionals who will be equipped to deal with issues concerning human genetics policy development and implementation, particularly in the current environment of health-care reform and managed care. Graduates of this program will serve as a resource for consultation in the development of public-health policy concerning genetic disease.

Fellowship candidates must demonstrate exceptional basic understanding of and competence in human genetics; hold an earned degree in genetics, biology, life sciences, or a similar field; have a well-grounded and appropriately documented scientific and technical background; have a broad professional background in the practice of human genetics as demonstrated by national or international reputation; be cognizant of related nonscientific matters that impact on human genetics; exhibit sensitivity toward political and social issues; have a strong interest and some experience in applying personal knowledge toward the solution of social problems; be a member of ASHG; be articulate, literate, adaptable, and interested in working on long-range public policy problems; be able to work with a variety of people of diverse professional backgrounds; and function well during periods of intense pressure.

The first fellow is working in the office of Senator Wellstone, Democrat from Minnesota, and devoting most of his time to studying and commenting on health-care and science issues.

DOE Grant No. DE-FG02-95ER61974.

A Hispanic Educational Program for Scientific, Ethical, Legal, and Social Aspects of the Human Genome Project

Margaret C. Jefferson and **Mary Ann Sesma**[1]
Department of Biology and Microbiology; California State University; Los Angeles CA 90032
213/343-2059, Fax: -2095, *mjeffer@flytrap.calstatela.edu*
http://vflylab.calstatela.edu/hgp
[1]Los Angeles Unified School District

The primary objectives of this grant are to develop, implement, and distribute culturally competent, linguistically appropriate, and relevant curriculum that leads to Hispanic student and family interactions regarding the science, ethical, legal, and social issues of the Human Genome Project. By opening up channels of familial dialogue between parents and their high school students, entire families can be exposed to genetic health and educational information and opportunities. In addition, greater interaction is anticipated between students and teachers, and parents and teachers. In the Los Angeles Unified School District alone, over 65% of the approximately 850,000 student enrollment are bilingual Hispanics. The 1990 census data revealed that the U.S.A. had a total population of 248,709,873, of which 22,354,059 were Hispanics, and thus, there is a need for materials to be disseminated throughout the U.S.A. that are relevant and understandable to this population.

Student curriculum consists of BSCS HGP-ELSI curriculum available in both English and Spanish; supplemental lesson plans developed and utilized by high school teachers in predominantly Hispanic classrooms that will be available via the World Wide Web; student-developed surveys that ascertain knowledge and perceptions of genetics and HGP-ELSI in Hispanic and other ethnic communities in the greater Los Angeles area; the University of Washington High School Human Genome Program exercises on DNA synthesis and sequencing; and career ladders and opportunities in genetics. The supplemental lesson plans are focused on four major units: the Cell; Mendelian Genetics and its Extensions; Molecular Genetics; and the Human Genome Project and ELSI. The concise concepts underlying each unit are being utilized in two ways: (a) first, the student activities emphasize logical, problem-solving exercises; tools or technologies applicable to that concept; when and where appropriate, a focus on the Hispanic population; and an understanding of the problems and compassion for the families associated with learning of genetic diseases. (b) second, the concepts serve as the springboard for the topics that the students include in science newsletters to their parents. In addition to on-campus activities, we intend to arrange field trips and/or classroom demonstrations of genetic and molecular biology techniques by scientists and other experts. The speakers would also be asked to discuss career opportunities and the educational requirements needed to enter the specific careers presented.

The parent curriculum consists of two major activities. First the student-parent newsletter is designed to drawn the parents into the curriculum. Students write newsletters on a biweekly basis. Each newsletter relates to a student curriculum subunit and the specific subunit concepts. English, Spanish, social science as well as biology and chemistry teachers assist the students in its production. The other major activity that involves the parents are the parent focus groups. Parents from each participating school are invited to monthly focus groups at their specific campus. The focus groups discuss issues related to genetics and health, legal and social issues as well as science issues that stem from the student newsletters. The discussions are in both English and Spanish with translators available. Links with other programs have been established.

DOE Grant No. DE-FG03-94ER61797.

Implications of the Geneticization of Health Care for Primary Care Practitioners

Mary B. Mahowald, John Lantos, Mira Lessick, Robert Moss, Lainie Friedman Ross, Greg Sachs, and Marion Verp
Department of Obstetrics and Gynecology and MacLean Center for Clinical Medical Ethics; University of Chicago; Chicago, IL 60637
312/702-9300, Fax: -0840, *mm46@midway.uchicago.edu*
http://ccme-mac4.bsd.uchicago.edu/CCMEHomePage.html

"Geneticization" refers to the process by which advances in genetic research are increasingly applicable to all areas of health care.[1] Studies show that primary caregivers are often deficient in their knowledge of genetics and genetic tests, and the ethical, legal, and social implications of this knowledge.[2-6] Accordingly, this project prepares primary caregivers who have no special training in genetics or genetic counseling to deal with the implications of the Human Genome Project for their practice.

Phase I (fall 1995): Generic topics will be addressed by PI and Co-PIs with Robert Wood Johnson clinical scholars and clinical ethics fellows, led by visiting or internal experts.

Topics: Goals, Methods, & Achievements of the HGP; Typology of Genetic Conditions; Scientific, Clinical, Ethical, and Legal Aspects of Gene Therapy; Concepts of Disease; Genetic Disabilities; Gender and Socio-economic Differences; Cultural and Ethnic Differences; Directive or Nondirective genetic counseling.

Speakers: Jeff Leiden; Julie Palmer; Dan Brock; Anita Silvers; Abby Lippman; James Bowman; Beth Fine

Phase II (Jan.–Mar. 1996): Teams of individuals, all trained in the same area of primary care, will identify and address issues specific to their area, developing course outlines, bibliography, and methodology based on grand rounds given by national expert.

Primary Care Area
Pediatrics: Genetics expert: Stephen Friend, Ethics Expert: Lainie F. Ross + fellow
Obstetrics/Gynecology: Genetics expert: Joe Leigh Simpson, Ethics Expert: Marion Verp + fellow
Medicine: Genetics expert: Tom Caskey, Ethics Expert: Greg Sachs + fellow
Family medicine: Genetics expert: Noralane Lindor, Ethics Expert: Robert Moss + fellow
Nursing: Genetics expert: Mira Lessick, Ethics Expert: Colleen Scanlon + fellow

Phase III (Apr.–May 1996): Policy issues will be identified and addressed as above for all areas of primary care, based on grand rounds given by national expert.

Policy team: Genetics expert: Sherman Elias; Ethics expert: John Lantos + trainee

Phase IV (Oct.–Dec. 1996): Presentation of content developed to new group of fellows and scholars by each of the above teams, followed by evaluation & revision.

Phase V (spring 1997): NATIONAL CONFERENCE and CME/CNE WORKSHOPS for primary caregivers, keynoted by Victor McKusick.

DOE Grant No. DE-FG02-95ER61990.

References

[1] Lippman A., Prenatal genetic testing and screening, *Amer J Law & Med* XVII, 15-50 (1991).

[2] Hofman, K.J., Tambor, E.S., Chase, G.A., Geller, G., Faden, R.R., and Holtzman, N.A., Physicians' knowledge of genetics and genetic tests, *Acad Med* **68**, 625-32 (1993).

[3] Holtzman, N.A., The paradoxical effect of medical training, *J Clin Ethics* **2**, 24142 (1992).

[4] Forsman, I, Education of nurses in genetics, *Amer J of Hum Genetics* 552-58, (1988).

[5] Williams, J.D., Pediatric nurse practitioners' knowledge of genetic disease *Ped Nursing* **9**, 1 19-21 (1983).

[6] George, J.B., Genetics: Challenges for nursing education, *J Ped Nursing* **7**, 5-8, (1992).

Nontraditional Inheritance: Genetics and the Nature of Science; Instructional Materials for High School Biology

Joseph D. McInerney and **B. Ellen Friedman**
Biological Sciences Curriculum Study; Colorado Springs, CO 80918
719/531-5550, Fax: -9104, *jmcinerney@cc.colorado.edu*

There often is a gap between the public's and scientists' views of new research findings, particularly if the public's understanding of the nature of science is not sound. Large quantities of new evidence and consequent changes in scientific explanations, such as those associated with the Human Genome Project and related genetics research, can accentuate those different views. Yet an appealing secondary effect of the unusually fast acquisition of data is that our view of genetics is changing rapidly during a brief time period, a relatively recent phenomenon in the field of biological sciences. This situation provides an outstanding opportunity to communicate the nature and methods of science to teachers and students, and indirectly to the public at large. The immediacy of new explanations of genetic mechanisms lets nontechnical audiences actually experience a changing view of various aspects of genetics, and in so doing, gain an appreciation of the nature of science that rarely is felt outside of the research laboratory.

The Biological Sciences Curriculum Study (BSCS) is developing a curriculum module that brings this active view of the nature and methods of science into the classroom via examples from recent discoveries in genetics. We will distribute this print module free of charge to interested high school biology teachers in the United States.

The examples selected for classroom activities include the instability of trinucleotide repeats as an explanation of genetic anticipation in Huntington disease and myotonic dystrophy, and the more widespread genetic mechanism of extranuclear inheritance, illustrated by mitochondrial inheritance. Background materials for teachers discuss a wider range of phenomena that require nontraditional views of inheritance, including RNA editing, genomic imprinting, transposable elements, and uniparental disomy. The genetics topics in the module share the common characteristic that they are not adequately explained by the traditional, Mendelian concepts that are taught in introductory biology at the high school level. In addition to updating the genetics curriculum and communicating the nature of science, the module devotes one activity to the ethical and social aspects of new genetics discoveries by challenging students to consider the current reluctance to test asymptomatic minors for the presence of the HD gene.

The major challenge we have faced in this project is to make relatively technical genetics information accessible to high school teachers and students and to turn the often passive treatment of scientific processes into an active experience that helps students develop an understanding and appreciation of the nature and methods of science. The module is being field tested in classrooms across the country. Evaluation data from the field test will guide final revision of the module prior to distribution.

DOE Grant No. DE-FG03-95ER61989.

The Human Genome Project: Biology, Computers, and Privacy: Development of Educational Materials for High School Biology

Joseph D. McInerney, **Lynda B. Micikas**, and B. Ellen Friedman
Biological Sciences Curriculum Study; Colorado Springs, CO 80918
719/531-5550, Fax: -9104, *jmcinerney@cc.colorado.edu*

One of the challenges faced by the Human Genome Project (HGP) is to handle effectively the enormous quantities and types of data that emerge as a result of progress in the project. The informatics aspect of the HGP offers an excellent example of the interdependence of science and technology. In addition, the electronic storage of genomic information raises important questions of ethics and public policy, many revolving around privacy.

The Biological Sciences Curriculum Study (BSCS) addresses the scientific, technological, ethical, and policy aspects of genome informatics in the instructional program titled The *Human Genome Project: Biology, Computers, and Privacy*. The program, intended for use in high school and college biology, consists of software and a 150-page print module. The software includes two model databases: a research database housing anonymous data (map data, sequence data, and biological/clinical information) and a registry that attaches names of 52 fictitious individuals (three kindreds) to genomic data. Students manipulate the database software as they work through seven classroom inquiries described in the print material. Also included is 50 pages of background material for teachers.

An introductory activity lets students become familiar with the software and dramatically demonstrates the advantages of technology in analysis of sequence data. In activities 1 and 2, students use the database to construct pedigrees and make initial choices about privacy with regard to genetic tests for their fictitious person. Activity 3 expands genetic anticipation, and in activities 4 and 5, students deal in depth with decision-making, ethics, and public policy, revisiting their earlier decision about testing and data accessibility. A final extension activity shows how comparisons with genomic data can be used to test hypotheses about the biological relationships between individual humans and about the evolutionary significance of DNA sequence similarities between different species.

External reviews and evaluation data from a field test involving 1,000 students in schools across the United States were used to guide final revision of the materials. BSCS will distribute the module free of charge to more than 10,000 high school and college biology teachers.

DOE Grant No. DE-FG03-93ER61584.

Involvement of High School Students in Sequencing the Human Genome

Maureen M. Munn, **Maynard V. Olson**, and **Leroy Hood**
Department of Molecular Biotechnology; University of Washington; Seattle, WA 98195
206/616-4538, Fax: /685-7344, *mmunn@u.washington.edu*

For the past two years, we have been developing a program that involves high school students in the excitement of genetic research by enabling them to participate in sequencing the human genome. This program provides high school teachers with the proper training, equipment, and support to lead their students through the exercise of sequencing small portions of DNA. The participating classrooms carry out two experimental modules, DNA synthesis (an introduction to DNA replication and the techniques used to study it) and DNA sequencing. Both of these experiments consist of three parts–synthesizing DNA fragments using Sequenase and a biotinlabeled primer, bench top electrophoresis using denaturing polyacrylamide gels, and colorimetric DNA detection that is specific for the biotinylated primer. Students analyze their sequencing data and enter it into a DNA assembly program. This year, in collaboration with Eric Lynch and Mary-Claire King from the Department of Genetics at the University of Washington, the students will be sequencing a region of chromosome 5q that may be involved in a form of hereditary deafness.

Students also consider the ethical, legal and social issues (ELSI) of genome research in a unit that explores the topic of presymptomatic testing for Huntington's disease (HD). This module was developed by Sharon Durfy and Robert Hansen from the Department of Medical History and Ethics at the University of Washington. It provides a scenario about a family that carries the HD allele, descriptions of the clinical and genetic aspects of the disorder, an exercise in drawing pedigrees and an autoradiograph showing the PCR assay used to detect HD. Students use an ethical decision-making model to decide whether, as a character from the scenario, they would be tested presymptomatically for the HD allele. Through this experience, they develop the skills to define ethical issues, ask and research the relevant questions about a particular topic and make justifiable ethical decisions.

In the first two years of this program, our focus was on the development of robust, classroom friendly modules that can be presented in up to six classes at one time. This year we will focus on disseminating this program to local, regional, and national sites. During a week-long workshop in July, 1995, we trained an additional thirteen high school teachers, bringing our current number to twenty teachers at thirteen schools. We have recruited local scientists to act as mentors to each of the schools and provide classroom support. On the regional level, four of our teachers are from outside the greater Seattle area and will be supported during the classroom experiments by scientists in their region. We have presented this program at national meetings and workshops, including the Human Genome Teacher Networking Project Workshop in Kansas City, KS (June, 1995) and the meeting of the National Association of Biology Teachers in Phoenix, AZ (October 1995). We have also distributed our modules to teachers and scientists throughout the nation to encourage the development of similar programs. This year we will also develop and pilot a module using automated sequencing. This will enable distant schools to participate in the program by providing them with the option of sending their DNA samples to the UW genome center for electrophoresis .

While we hope the human genome sequencing experience will interest some students in science careers, a broader goal is to encourage high school students to think constructively and creatively about the implications of scientific findings so that the coming generation of adults will make judicious decisions affecting public policies.

DOE Grant No. DE-FG03-96ER62175.

The Gene Letter: A Newsletter on Ethical, Legal, and Social Issues in Genetics for Interested Professionals and Consumers

Philip J. Reilly, **Dorothy C. Wertz**, and **Robin J.R. Blatt**[1]
The Shriver Center for Mental Retardation; Division of Social Science, Ethics and Law; Waltham, MA 02254
617/642-0230, Fax: /893-5340, *preilly@shriver.org*
[1]Also at Massachusetts Department of Public Health, Boston, MA
http://www.shriver.org

We propose to develop a newsletter on ELSI-related issues for dissemination to a broad general audience of professionals and consumers. No such focussed public newsletter currently exists. Entitled *The Gene Letter*, the newsletter will be distributed monthly on-line, through the Internet. Updated weekly on the Internet, it will be poised to react in a timely fashion to new developments in science, law, medicine, ethics, and culture. The newsletter does not propose to provide comprehensive education in genetics for the American public, but rather to begin an information network that interested people can use for further information. It will be the most widely-distributed newsletter on ELSI genetics in the world, with the largest consumer readership. Features will be largely informational and will include new scientific/medical developments and attendant ELSI issues, new court decisions, legislation, and regulations, balanced responses to new concerns in the media, and new developments related to health that may be of interest to health care providers and consumers. Features will present balanced opinions. An editorial board will review each issue, prior to publication, for cultural sensitivity, emphasis, balance, and concerns of persons with disabilities. *The Gene Letter* will also include factual information on upcoming events, new ELSI research, where to find genetics on the Internet, new publications (annotated), and where to find further information about each feature. Readers will be invited to send letters, queries, news, bibliography, comments, and consumer concerns either on *The Gene Letter* Internet chatroom or in hard copy. A hard copy of the first on-line issue will be used to assess readers' needs and interests. It will be distributed to 500 community college students representing blue-collar ethnic groups, and to 2000 members of a broad general audience.

A special evaluation of readers' knowledge and ethical/ social concerns raised by *The Gene Letter* will take place at the end of the second year in order to assess outcome. It is our intention that *The Gene Letter* become self-supporting after two years.

DOE Grant No. DE-FG02-96ER62174.

The DNA Files: A Nationally Syndicated Series of Radio Programs on the Social Implications of Human Genome Research and Its Applications

Bari Scott, Matt Binder, and Jude Thilman
Genome Radio Project; KPFA-FM; Berkeley, CA 94704
510/848-6767 ext 235, Fax: /883-0311, *strp@aol.com*

The DNA Files is a series of nationally distributed public radio programs furthering public education on developments in genetic science. Program content is guided by a distinguished body of advisors and will include the voices of prominent genetic researchers, people affected by advances in the clinical application of genetic medicine, members of the biotech industry, and others from related fields. They will provide real-life examples of the complex social and ethical issues associated with new discoveries in genetics. In addition to the general public radio audience, the series will target educators, scientists, and involved professionals. Ancillary educational materials will be distributed in paper and digital form through over two dozen collaborative organizations and fulfillment of listener requests.

"DNA and Behavior: Is Our Fate Written in Our Genes?" is the pilot documentary for the series, scheduled for release in early 1996. The show will help the lay person understand and evaluate recent research in the area of behavioral genetics. Recently, we've seen news media reports on newly discovered genetic factors being related to behaviors such as alcoholism, mental illness, sexual orientation and aggression. This program will look at several examples of these "genetic factors" and evaluate the strengths and weaknesses of various methodologies involved in the research; and introduce such controversial issues as the re-emergence of a eugenics movement based on theoretical suppositions drawn from recent work in behavioral genetics.

With information linking major diseases such as breast cancer, colon cancer, and arteriosclerosis to genetic factors, new dangers in public perception emerge. Many people who hear about them mistakenly conclude that these diseases can now be easily diagnosed and even cured. On the other end of the public perception spectrum, unfounded fears of extreme, and highly unlikely, consequences also appear. Will society now genetically engineer whole generations of people with "designer genes" offering more "desirable physical qualities"? The *DNA Files* will ground public understanding of these issues in reality. "DNA and the Law" reviews the scientific basis for genetic fingerprinting and looks at cases of alleged genetic discrimination by insurance companies, employers and others. This program also looks at disputes over paternity, intellectual property rights, the commercialization of genetic information, informed consent and privacy issues. Other shows include "The Search for a Breast Cancer Gene," "Prenatal Genetic Testing and Treatment," "Evolution and Genetic Diversity," "Sickle-Cell Disease and Thalassemia: Hope for a Cure," and "Theology, Mythology and Human Genetic Research."

DOE Grant No. DE-FG03-95ER62003.

Communicating Science in Plain Language: The Science+ Literacy for Health: Human Genome Project

Maria Sosa, Judy Kass, and Tracy Gath
American Association for the Advancement of Science; Washington, DC 20005
202/326-6453, Fax: /371-9849, *msosa@aaas.org*

Recent literacy surveys have found that a large number of adults lack the skills to bring meaning to much of what is written about science. This, in effect, denies them access to vital information about their health and well-being. To address this need, the American Association for the Advancement of Science (AAAS) is developing a 2-year project to provide low-literate adults with the background knowledge necessary to address the social, ethical, and legal implications of the Human Genome Project.

With its **Science + Literacy for Health: Human Genome Project**, AAAS is using its existing network of adult education providers and volunteer science and health professionals to pursue the following overall objectives: (1) to develop new materials for adult literacy classes, including a high-interest reading book and accompanying curriculum, an implementation framework, a short video providing background information on genetics, a database of resources, and fact sheets that will assist other organizations and researchers in preparing easy-to-read materials about the human genome project, and (2) to develop and conduct a campaign to disseminate project materials to libraries and community organizations carrying out literacy programs throughout the United States.

Because not every low-literate adult is enrolled in a literacy class, our model for helping scientists communicate in simple language will have impact beyond classrooms and learning centers. In preliminary contacts, community groups providing health services have indicated that the proposed materials are not only desirable but needed; indeed such groups often receive requests for information on heredity and genetics. The module developed by AAAS should enable other medical and scientific organizations to communicate more effectively with economically disadvantaged populations, which often include a large number of low-literate individuals.

DOE Grant No. DE-FG02-95ER61988.

The Community College Initiative

Sylvia J. Spengler and Laurel Egenberger
Lawrence Berkeley National Laboratory; Berkeley, CA 94720
510/486-4879, Fax: -5717, *sjspengler@lbl.gov*
http://csee.lbl.gov/cup/ccibiotech/Index.html

The Community College Initiative prepares community college students for work in biotechnology. A combined effort of Lawrence Berkeley National Laboratory (LBNL) and the California Community Colleges, we aim to develop mechanisms to encourage students to pursue science studies, to participate in forefront laboratory research, and to gain work experience. The initiative is structured to upgrade the skills of students and their instructors through four components.

Summer Student Workshops: Four weeks summer residential programs for students who have completed the first year of the biotechnology academic program. Ethical, legal and social concerns are integrated into the laboratory exercises and students learn to identify commonly shared values of the scientific community as well as increase their understanding of issues of personal and public concern.

Teacher Workshop Training: Seminars for biotechnology instructors to improve, upgrade, and update their understanding of current technology and laboratory practices, with emphasis on curriculum development in current topics in ethical, legal, and social issues in science.

Sabbatical Fellowships: For community college instructors to provide investigative and field experience in research laboratories. During the fellowship, teachers also assist in development of student summer research activities.

Summer Faculty-Student Teams: Post-fellowship faculty and biotechnology students who have finished their second year of study team on a research project.

Genome Educators

Sylvia Spengler and **Janice Mann**
Human Genome Program; Life Sciences Division; Lawrence Berkeley National Laboratory; Berkeley, CA 94720
510/486-4879, Fax: -5717, *sjspengler@lbl.gov* or *jlmann@lbl.gov*
http://www.lbl.gov/Education/Genome

Genome Educators is an informal network of educational professionals who have an active interest in all aspects of genetics research and education. This national group includes scientists, researchers, educational curriculum developers, ethicists, health professionals, high school teachers and instructors at college and graduate levels, and others in occupations affected by genetic research.

Genome Educators is a unique collaborative effort dedicated to sharing information and resources to further understanding of current advances in the field of genetics. Seminars, workshops, and special events are sponsored at frequent intervals. Genome Educators maintains an active World Wide Web site (URL: http://www.lbl.gov/Education/Genome). This site contains a calendar of events, directory of participating genome educators, and information about educational resources and reference tools. Participating genome educators may publish articles and talks of interest at this site. In addition, a monitored discussion group is maintained to facilitate dialog and resource sharing among participants.

Getting the Word Out on the Human Genome Project: A Course for Physicians

Sara L. Tobin and **Ann Boughton**[1]
Department of Biochemistry and Molecular Biology; Center for Biomedical Ethics; Stanford University; Palo Alto, CA 94304-1709
415/725-2663, Fax: -6131, *tobinsl@leland.stanford.edu*
[1]Thumbnail Graphics; Oklahoma City, OK 73118

Progressive identification of new genes and implications for medical treatment of genetic diseases appear almost daily in the scientific and medical literature, as well as in public media reports. However, most individuals do not understand the power or the promise of the current explosion in knowledge of the human genome. This is also true of physicians, most of whom completed their medical training prior to the application of recombinant DNA technology to medical diagnosis and treatment. This lack of training prevents physicians from appreciating many of the recent advances in molecular genetics and may delay their acceptance of new treatment regimens. In particular, physicians practicing in rural communities are often limited in their access to resources that would bring them into the mainstream of current molecular developments. This project is designed to fill two important functions: first, to provide solid training for physicians in the field of molecular medical genetics, including the impact, implications, and potential of this field for the treatment of human disease; second, to utilize physicians as informed community resources who can educate both their patients and community groups about the new genetics.

We propose to develop a flexible, user-friendly, interactive multimedia CD-ROM designed for continuing education of physicians in applications of molecular medical genetics. To initiate these objectives, we will develop the design of the CD and will produce a prototype providing a detailed presentation of one of the four training areas. These areas are (1) Genetics, including DNA as a molecular blueprint, chromosomes as vehicles for genetic information, and patterns of inheritance; (2) Recombinant techniques, stressing cloning and analytical tools and techniques applied to medical case studies; (3) Current and future clinical applications, encompassing the human genome project, technical advances, and disease diagnosis and prognosis; and (4) Societal implications, focusing on approaches to patient counseling, genetic dilemmas faced by patients and practitioners, and societal values and development of an ethical consensus. Area (2) will be presented in the prototype.

The CD format will permit the use of animation, video, and audio, in addition to graphic illustrations and photographs. We will build on our existing base of computer generated illustrations. A hypertext glossary, user notes, practice tests, and customized settings will be utilized to tailor the CD to the needs of the user. Brief, multiple-choice examinations will be evaluated for continuing medical education credits by the Office of Continuing Medical Education. The CD will be programmed to permit updates of scientific and medical advances either by downloading from the Internet or from a disc available by subscription.

This is a cooperative project involving individuals with documented expertise in teaching of molecular medical genetics, continuing medical education, graphic design, and CD-ROM production. The content of the CD will be supervised by a scientific board of directors. We present mechanisms for the evaluation of the CD by rural Oklahoma physicians. Arrangements have been made for distribution of the CD by a national publisher of medical and scientific materials. This CD will provide a powerful tool to educate physicians and the public about the power and potential of the human genome project for the benefit of human health.

DOE Grant No. DE-FG03-96ER62172.

The Genetics Adjudication Resource Project

Franklin M. Zweig
Einstein Institute for Science, Health, and the Courts;
Bethesda, MD 20814
301/961-1949, Fax: /913-0448, *einshac@aol.com*
http://www.ornl.gov/courts

The Einstein Institute for Science, Health, and the Courts is preparing the foundation for a new utility needed to prepare the nation's 21,000 courts to adjudicate the genetics and ELSI-related issues that foreseeably will rush into the courtroom as the Human Genome Project completes its genomic mapping and sequencing mission during the next ten years. This project initiates practical collaboration among courts, legal and policy-making institutions, and science centers leading to modalities for understanding the scientific validity of claims, and for the resolution of ethical, legal, and social disputes arising within the genetic testing and gene therapy contexts. Our objective over the ensuing decade is to facilitate genetic testing and gene therapy dispute management, and to avoid to the extent possible the confusion that characterized adjudication of forensic DNA technologies during the decade just ended.

The outlines of a genetics adjudication utility were given form by the 1995 Working Conversation on Genetics, Evolution, and the Courts, involving 37 federal and state judges and others in science and policymaking leadership positions from across the nation. The courts are becoming aware of genetics, molecular biology, and their applications, and judges want public confidence to be maintained as the profound and complex issues set in motion by the HGP begin the long course of litigation. Modalities for understanding the underpinning science are needed, as well as instrumentalities to assure that the best cases are actually filed and pursued. Because the courts are the front-line for resolving disputes, creative lawyering will assure an abundance of lawsuits. Many such lawsuits will request the courts to make policy judgments, perhaps best undertaken by state legislatures and Congress. Accordingly, a new adjudication utility should provide forums for judicial/legislative exchange, preparatory deliberations in anticipation of pressure to make rushed policies under conditions of great social uncertainty in the wake of human genetics progress.

EINSHAC will provide a design, planning, communications, and implementation center for a multipurpose resource project available to the courts. It will undertake over an 18 month period the following tasks, pilot-testing each and assessing the best organizational locales for those that exhibit promise:

1. Judicial Education in Genetics & ELSI-Related Issues for six Judicial Branch leadership associations and nine metropolitan courts—aimed at 1,000 judges—in conjunction with scientific faculty and coaches mobilized by DOE/national laboratories and the American Society for Human Genetics.

2. Judicial Digital Electronic Collegium—technological modernization of the courts community by providing access to ELSI and genetics information through Internet resources.

3. Amicus Brief Development Trust Fund—a process and resources to support law development at the state and federal appeals courts level.

4. Genetics Indigent Party Trust Fund—a process and resources at the state and federal trial level to sustain meritorious civil cases holding promise of effective law development.

5. Establishment of a Pro-Bono Legal Services Clearinghouse—a personal and on-line referral resource for persons seeking representation for genetics and ELSI-related cases.

6. Access to Neutral Expert Witnesses—advisors to courts encountering particularly complex cases deemed right for the judicial exercise of Federal Rule of Evidence 706 and its State counterparts.

7. Pilot of Judicial/Legislative ELSI Policy Forums—provision of neutral staff and coordination in three mid-Atlantic states considering legislation related to health care, insurance, privacy, medical records.

8. National Training Center for Minority Justice Personnel—facilitating a leadership preparation program for the nation's minority court-related personnel in a consortium arrangement with the Ruffin Society of Massachusetts, the College of Criminal Justice at Northeastern University, and the Flaschner Judicial Institute.

The Project actively involves judges, scientists, and prominent lawyers. It will report to the *EINSHAC* Board of Directors that includes prominent judges, justices and scientists, several of whom participated in the 1995 Working Conversation on Genetics, Evolution and the Courts. As a continuing guidance forum, *EINSHAC* will conduct a Working Conversation followup in Orleans, Cape Cod in July, 1996.

DOE Grant No. DE-FG02-96ER62081.

Alexander Hollaender Distinguished Postdoctoral Fellowships

Linda Holmes and **Eugene Spejewski**
Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
423/576-3192, Fax: /241-5220, *holmesl@orau.gov* or *alexpgm@orau.gov*
http://www.orau.gov/oher/hollaend.htm

The Alexander Hollaender Distinguished Postdoctoral Fellowships, sponsored by the Department of Energy (DOE), Office of Health and Environmental Research (OHER), support research in the fields of life, biomedical, and environmental sciences. Since the DOE Human Genome Distinguished Postdoctoral Fellowships and DOE Global Change Distinguished Postdoctoral Fellowships both had their last application cycles in FY 1995, the Hollaender program is now open to recent PhD graduates in the fields of human genome and global change, as well.

Fellowships of up to 2 years are tenable at any DOE, university, or private laboratory providing the proposed adviser at that laboratory receives at least $150,000 per year in support from OHER. Fellows earn stipends of $37,500 the first year and $40,500 the second. To be eligible, applicants must be U.S. citizens or permanent residents at the time of application, and must have received their doctoral degrees within two years of the earliest possible starting date, which is May 1 of the appointment year.

The Oak Ridge Institute for Science and Education (ORISE), administrator of the fellowships, prepares and distributes program literature to universities and laboratories across the country, accepts applications, convenes a panel to make award recommendations, and issues stipend checks to fellows. The review panel identifies finalists from which DOE selects the award winners. Deadline for the FY 1999 fellowship cycle is January 15, 1998. For more information or an application packet, contact Linda Holmes at the Oak Ridge Institute for Science and Education, P. O. Box 117, Oak Ridge, TN 37831-0117 (423/576-9975, Fax: /241-5220).

DOE Contract No. DE-AC05-760R00033.

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, Sheryl A. Martin, **John S. Wassom**, Judy M. Wyrick, Laura N. Yust, Murray Browne, and Marissa D. Mills
Life Sciences Division; Oak Ridge National Laboratory; Oak Ridge, TN 37830
423/576-6669, Fax: /574-9888, *bkq@ornl.gov*
http://www.ornl.gov/hgmis

The Human Genome Management Information System (HGMIS), established in 1989, provides information about the international Human Genome Project in print and World Wide Web formats to both technical and general audiences. HGMIS is sponsored by the Human Genome Program Task Group of the DOE Office of Biological and Environmental Research to help fulfill DOE's commitment to informing scientists, policymakers, and the public about the program's funded research and the context in which the research is conducted. Several HGMIS products, including the Web sites and newsletter, have won technical and electronic communication awards.

HGMIS goals center on facilitating research at the interface of genomics and other biological disciplines that seek revolutionary solutions to biological, environmental, and biomedical challenges. By communicating information about the Human Genome Project and its impact, HGMIS increases the use of project-generated resources, reduces duplicative research efforts, and fosters collaborations and contributions to biology from other research disciplines.

Furthermore, communicating scientific and societal issues to nonscientist audiences contributes to increased science literacy, thus laying a foundation for more informed decision making and public-policy development. For example, since 1995 HGMIS has been participating in a project to educate the judiciary about the basics of genetics and gene testing. The aim is to prepare judges for the flood of cases involving genetic evidence that soon will enter the nation's courtrooms.

Information Resources

In keeping with its goals, HGMIS produces the following information resources in print and on the Web:

Human Genome News (HGN). A quarterly forum for interdisciplinary information exchange, *HGN* uniquely presents a broad spectrum of topics related to the Human Genome Project in a single publication. Articles feature topics that include project goals, progress, and direction; available resources; applications of project data and resources to provide a better understanding of biological processes; related or spinoff programs; medical uses of genome data; ethical, legal, and social considerations; legislative updates; other publications; meeting calendars; and funding information. Most *HGN* articles also contain sources of additional information. In May 1997, DOE acknowledged the newsletter's value by presenting an exceptional service award to *HGN*'s managing editor at a symposium celebrating 50 years of biological and environmental research.

Among 14,000 domestic and foreign *HGN* subscribers are genome and basic researchers at universities, national laboratories, nonprofit organizations, and industrial facilities; educators; industry representatives; legal personnel; ethicists; students; genetic counselors; medical professionals; science writers; and other interested individuals. All 41 issues of *HGN*, indexed and searchable, are accessible via the HGMIS Web site.

Other Publications. HGMIS also produces the DOE *Primer on Molecular Genetics*, progress reports on the DOE Human Genome Program, Santa Fe contractor-grantee workshop proceedings, 1-page topical handouts, and other related resource documents. Expanded and revised by HGMIS from an earlier DOE document, the DOE *Primer on Molecular Genetics* continues to be in demand. It is used as a handout for genome centers; a resource for new staff training by companies that make products for genome scientists; and an educational tool for teachers, genetic counselors, and such organizations as high schools, universities, and medical schools for student and continuing-education curricula. More than 35,000 hard copies have been distributed. The primer also is available in several formats at the HGMIS Web site, including an Adobe Acrobat version that can be used to print "originals" from users' printers.

Distribution of Documents. HGMIS has distributed more than 65,000 copies of items requested by subscribers, meeting attendees, and managers of genetics meetings and educational events. These items include HGN, program and workshop reports, DOE-NIH 5-year plans, DOE *Primer on Molecular Genetics*, and *To Know Ourselves*. On request, HGMIS supplies multiple copies of publications for meetings and educational purposes.

Electronic Communications. In November 1994, HGMIS began producing a comprehensive, text-based Web server called Human Genome Project Information, which is devoted to topics relating to the science and societal issues surrounding the genome project. In July 1997, this site was divided to better serve the two diverse audience categories that represent the majority of users: scientists and the public. The sites contain more than 1700 text files that are accessed over 1.2 million times each year. Each month, about 10,000 host computers connect to the HGMIS sites directly and through more than 1000 other Web sites. In addition, HGMIS links to the National Institutes of Health and international Human Genome Organisation sites, as well as to sites dedicated to education and to the ethical, legal, and social implications of the Human Genome Project.

All HGMIS publications are published on the Web site, along with such DOE-sponsored documents as *Your Genes, Your Choices*; the Genetic Privacy Act; and historical and other documents pertaining to the Human Genome Project. HGMIS collaborates with the Einstein Institute for Science, Health, and the Courts to produce *CASOLM*, the online magazine for judicial education in genetics and biomedical issues. HGMIS also maintains the Genetics section of the Virtual Library from CERN (Switzerland) and the DOE Human Genome Program pages and moderates the BioSci Human Genome Newsgroup.

Information Source

HGMIS answers individual questions and supplies general information about the Human Genome Project by telephone, fax, and e-mail and, as appropriate, links scientists with questions to appropriate Human Genome Project contacts. HGMIS staff exchange ideas and suggestions with investigators, industry representatives, and others when attending occasional scientific conferences and genome-related meetings and displaying the DOE Human Genome Project traveling exhibit. HGMIS staff also make presentations on the Human Genome Project to educational, judicial, and other groups.

HGMIS resources serve as a primary source for the popular media and for discipline-specific publications that broaden the distribution of genome project information by extracting and reprinting from HGMIS resources and by linking to various parts of the HGMIS Web site.

HGMIS continuously monitors changes in the direction of the international Human Genome Project and searches for ways to strengthen the content relevancy of the newsletter, the Web site, and other services.

DOE Contract No. DE-AC05-96OR22464.

Human Genome Program Coordination

Sylvia J. Spengler
Lawrence Berkeley National Laboratory; Berkeley CA 94720
510/486-4879, Fax: -5717, *sjspengler@lbl.gov*
http://www.lbl.gov/Education/ELSI

The DOE Human Genome Program of the Office of Health and Environmental Research (OHER) has developed a number of tools for management of the Program. Among these was the Human Genome Coordinating Committee (HGCC), established in 1988. In 1996, the HGCC was expanded to a broader vision of the role of genomic technologies in OHER programs, and the name was changed to reflect this broadening. The HGCC is now the Biotechnology Forum. The Forum is chaired by the Associate Director, OHER. Members of the Human Genome Program Management Task group are ex officio members, as are members of the Health and Environmental Research Advisory Committee's subcommittee on the Human Genome. Responsibilities of the Forum include: assisting OHER in overall coordination of DOE-funded genome research; facilitating the development and dissemination of novel genome technologies; recommending establishment of ad hoc task groups in specific areas, such as informatics, technologies, model organisms; and evaluation of progress and consideration of long-term goals. Members also serve on the Joint DOE-NIH Subcommittee on the Human genome, for interagency coordination. The coordination group also participates in interface programs with other facilities and provides scientific support for development of other OHER goals, as requested.

Support of Human Genome Program Proposal Reviews

Walter Williams
Education/Training Division; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
423/576-4811, Fax: /241-2727, *williamw@orau.gov*

The Oak Ridge Institute for Science and Education (ORISE), operated by Oak Ridge Associated Universities, provides assistance to the DOE Office of Health and Environmental Research in the technical review of proposals submitted in response to solicitations by the DOE Human Genome Program. ORISE staff members create and maintain a database of all proposal information; including abstracts, relevant names and addresses, and budget data. This information is compiled and presented to proposal reviewers. Before review meetings, ORISE staff members make appropriate hotel and meeting arrangements, provide each reviewer with proposal copies and evaluation guidelines, and coordinate reviewer travel and honoraria payment. Onsite meeting support includes collecting all reviewer evaluation forms and scores, entering reviewer scores into the database, preparing appropriate reports, providing onsite computer support, and handling all logistical issues. Other support includes assistance with program advertising and preparation of reviewer comments following each review. ORISE may also assist with pre- and post-review activities related to conferences, seminars, and site visits.

DOE Contract No. DE-AC05-76OR00033.

Former Soviet Union Office of Health and Environmental Research Program

James Wright
Education/Training Division; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
423/576-1716, Fax: /241-2727, *wrightj@orau.gov*

The Former Soviet Union Office of Health and Environmental Research Program, sponsored by the U.S. Department of Energy, Office of Health and Environmental Research, recognizes outstanding scientists in the field of health and environmental research from the independent states of the former Soviet Union. The program fosters the international exchange of new ideas and innovative approaches in health and environmental research; strengthens ties and encourages continuing collaboration among Russians and U.S. scientists; and establishes and maintains environmental research capability in the former Soviet Union. The program has supported more than 23 Russian principal investigators and approximately 110 other research associates in Moscow, St. Petersburg, and Novosibirsk. More importantly, the program has enabled many high quality Russian biological, genome informatics, physical mapping and mutagenesis detection, human genetics,, biochemistry, DNA sequencing technology, protein analysis, molecular genetics, and other related research infrastructures to continue operating in an uncertain economic environment.

DOE Contract No. DE-AC05-76OR00033.

1996 Phase I

An Engineered RNA/DNA Polymerase to Increase Speed and Economy of DNA Sequencing

Mark W. Knuth
Promega Corporation; Madison, WI 53711-5399
608/274-4330, Fax: /277-2601

DNA sequence information is the cornerstone for considerable experimental design and analysis in the biological sciences. The proposed studies will focus on advancing DNA sequencing by creating a new enzyme that eliminates the need for an oligonucleotide primer to initiate DNA synthesis at a defined site, and that can use dideoxy nucleotides for chain termination. The new method should reduce the time and cost required to obtain DNA sequences and enhance the speed and cost effectiveness of current DNA sequencing technologies. Phase I studies will focus on purifying mutant T7 RNA polymerases known to incorporate dNTPs into DNA chains, developing protocols for rapid small scale mutant enzyme purification, evaluating the purified mutants for properties relevant to DNA sequencing, developing facile mutagenesis schemes and producing mutant RNA/DNA polymerases with altered promoter recognition. The results from phase I will provide the foundation for Phase II research, which will focus on refining properties of the mutant by: (1) expanding the number of mutations examined using the purification protocols, assays, and mutagenesis screening methods developed in Phase I and (2) examining the effect of each mutation on enzymatic properties important to DNA sequencing applications, and (3) optimizing conditions for sequencing performance. In Phase III, Promega will commercialize the new mutant enzymes through its own extensive distribution network and by collaborating with major instrumentation firms to adapt the technology to automated DNA sequencing systems.

DOE Grant No. DE-FG02-96ER8226.

Directed Multiple DNA Sequencing and Expression Analysis by Hybridization

Gualberto Ruano
BIOS Laboratories, Inc.; New Haven, CT 06511
800/678-9487 or 203/773-1450, Fax: 800/315-7435 or 203/562-9377

The overall goal of this project is to develop molecular resources with direct applications to either DNA sequence analysis or gene expression analysis in multiplexed formats using sequential hybridization of Peptide Nucleic Acid (PNA) oligomer probes. PNA oligomers hybridize more stably and specifically to cognate DNA targets than conventional DNA oligonucleotides. The Phase I project discussed here is concerned with development of PNA probe technology having direct application either to the directed sequencing process or to gene expression profiling. With regard to directed sequencing, we seek improvements in the three multiply repeated steps associated with this process, namely (1) probe assembly, (2) sequencing reactions, and (3) gel electrophoresis. In PNA hybridization sequencing, sequences are generated directly from the template by multiplex DNA sequencing using anchor primers known to have frequent annealing sites. Electrophoresis is performed en masse for each anchor primer reaction, blotted to nylon membranes and individual sequences are selectively exposed by iterative hybridization to specific 8-mer PNA probes derived from sequences statistically over-represented in expressed DNA and obtained from a pre-synthesized library. Additionally, the same PNA library can be used as a source of hybridization probes for querying expression patterns of specific genes in any cell line or tissue. Specific gene expression can be monitored by coupling gene-specific RT-PCR with hybridization when cDNA products are separated by gel electrophoresis and blotted to nylon membranes. Patterns of gene expression are then resolved by hybridization using PNA oligomers. Bands corresponding to specific genes can be deconvoluted using sequence information from RT-PCR primers and PNA probes. Higher throughput expression analysis can be achieved by multiplexed gel electrophoresis, blotting and iterative probing of RT-PCR reactions with individual PNA probes.

DOE Grant No. DE-FG02-96ER8213.

1996 Phase II

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogeneous Public Genome Databases

Joseph Leone
CyberConnect Corporation; Storrs, CT 06268
860/486-2783, Fax: /429-2372

The interoperability of public genome databases is expected to be crucial in making the Human Genome Project a success. This project will develop software tools in which users in the genome community can learn or examine public genome database schemes in a relatively short time and can produce a correct Structured Query Language (SQL) expression easily. In Phase I, a concept system was constructed and the effectiveness of formulating ad hoc queries graphically was demonstrated. Phase II will focus on transforming the concept system into a product that is robust and portable. Two types of computer programs will be developed. One is a client program which is to be distributed to community users who intend to access public genomic databases and link them with local databases. The other is a server program and a suite of software tools designed to be used by those genome centers which intend to make their databases publicly accessible.

DOE Grant No. DE-FG02-95ER81906.

Low-Cost Automated Preparation of Plasmid, Cosmid, and Yeast DNA

Tuyen Nguyen, Randy F. Sivila, Joshua P. Dyer, and **William P. MacConnell**
MacConnell Research Corporation; San Diego, CA 92121
619/452-2603, Fax: -6753

MacConnell Research currently manufactures and sells a low cost automated bench-top instrument that can purify up to 24 samples of plasmid DNA simultaneously in one hour at a cost of $0.65 per sample and under $8000 for the instrument. The patented instrument uses a form of agarose gel electrophoresis to purify the plasmid DNA and electroelutes into approximately a 20 +l volume. The instrument has many advantages over other robotic and manual methods including the fact that is it two times faster, at least six times less expensive, much smaller in size, easier to operate, less cost per sample, and results in DNA pure enough for direct use in fluorescent automated sequencing. The instrument process begins with bacterial culture which is loaded directly into a disposable cassette in the machine.

In Phase II work we are developing an instrument which simultaneously purifies plasmid DNA from up to 192 (2 X 96) bacterial samples in 1.5 hours. Prototypes of this instrument thus far constructed have allowed the purification of 3–7 micrograms of high purity plasmid DNA per lane from 1.5 ml of bacterial culture. We have attempted to optimize all of the: instrument electrophoretic run parameters, lysis chemistry, lysis reagent delivery devices, reagent storage at room temperature, desalting processes and overall instrument mechanical and electronic control. Instrument prototypes have also been used to prepare cosmid or yeast DNA in quantities of 1–5 micrograms per cassette lane. Trials thus far have yielded plasmid DNA of sufficient purity for direct use in automated fluorescent and manual sequencing as well as other molecular biology protocols. We have studied the purity of the resulting DNA when directly sequenced on a Licor 4000 Long Reader and ABI 373A automated DNA sequencers. Results from the Licor 4000 instrument give routine read lengths of >850 base pairs with 98% accuracy while ABI 373A reads generally exceed 400 base pairs with similar accuracy.

The proposed 2 X 96-channel instrument will purify up to 1200 plasmid DNA preps per eight hour day. It will significantly reduce the cost and technician labor of high throughput plasmid DNA purification for automated sequencing and mapping.

DOE Grant No. DE-FG03-94ER81802/A000.

GRAIL-GenQuest: A Comprehensive Computational Framework for DNA Sequence Analysis

Ruth Ann Manning
ApoCom, Inc.; Oak Ridge, TN 37830
423/482-2500, Fax: /220-2030

Although DNA sequencing in the Human Genome Project is occurring fairly systematically, biotechnology companies have focused on sequencing regions thought to contain particular disease genes. The client-server DNA sequence analysis system GRAIL is the most accurate and widely used computer-based system for locating and characterizing genes in DNA sequences, but it is not accessible to many biotechnology environments. The GRAIL client software and graphical displays have been developed for high-end UNIX-based computer workstations. Such workstations are standard equipment in universities and large companies, but personal computers (PCs) and Macintosh computers are the prevalent technology within the biotechnology community. This Phase I project will design Macintosh- and Windows-based client graphical user interface prototypes for GRAIL.

The growth of DNA databases is expected to continue at a fast pace in the attempt to sequence the human genome completely by the year 2005. Parallel processing is a viable solution to handle searching through the ever-increasing volume of data. During Phase I, genQuest—the sequence comparison server portion of the GRAIL system—will be parallelized for shared-memory platforms and will use PVM[1] for the development of genQuest servers on networks of PCs and workstations and other innovative, high-performance computer architectures.

Prototype graphical interface systems for Macintosh, NT Windows, and Windows 95 that mimic the function and operation of the current GRAIL-genQuest clients will enable a larger portion of biotechnology companies to make use of the GRAIL suite of analysis tools. Parallel genQuest servers will improve response time for searches and increase user capacity per server. Such fast shared- and distributed-memory computing solutions will improve the cost-performance ratio and make parallel searches more affordable to the biotechnology community using general multipurpose hardware.

DOE Grant No. DE-FG02-95ER81923.

[1] The Parallel Virtual Machine (PVM) message-passing library allows a collection of UNIX-based computers to function as a single multiple-processor supercomputer.

Projects Completed FY 1994–95

Projects in this section have been completed or did not receive support through the DOE Human Genome Program in FY 1996.

Sequencing

Sequencing by Hybridization: Methods to Generate Large Arrays of Oligonucleotides
Thomas M. Brennan

Sequencing by Hybridization: Development of an Efficient Large-Scale Methodology
Radomir Crkvenjakov

Genomic Instrumentation Development: Detection Systems for Film and High-Speed Gel-Less Methods
Jack B. Davidson and Robert S. Foote

Single-Molecule Detection Using Charge-Coupled Device Array Technology
M. Bonner Denton, Richard Keller, Mark E. Baker, Colin W. Earle, and David A. Radspinner

Coupling Sequencing by Hybridization with Gel Sequencing for Inexpensive Analysis of Genes and Genomes
Radoje Drmanac, Snezana Drmanac, and Ivan Labat

Physical Structure and DNA Sequence of Human Chromosomes
Glen A. Evans

Using Scanning Tunneling Microscopy to Sequence the Human Genome
Thomas L. Ferrell, Robert J. Warmack, David P. Allison, K. Bruce Jacobson, Gilbert M. Brown, and Thomas G. Thundat

DNA Sequence Analysis by Solid-Phase Hybridization
Robert S. Foote, Richard A. Sachleben, and K. Bruce Jacobson

DNA Sequencing Using Stable Isotopes
K. Bruce Jacobson, Heinrich F. Arlinghaus, Gilbert M. Brown, Robert S. Foote, Frank W. Larimer, Richard A. Sachleben, Norbert Thonnard, and Richard P. Woychik

Preparation of Oligonucleotide Arrays for Hybridization Studies
Michael C. Pirrung, Steven W. Shuey, David C. Lever, Lara Fallon, J.-C. Bradley, and William P. Hawe

Improvement and Automation of Ligation-Mediated Genomic Sequencing
Arthur D. Riggs and Gerd P. Pfeifer

*Analysis of a 53-Kb Nucleotide Sequence from the Right Genome Terminus of the Variola Major Virus Strain India-1967
Sergei N. Shchelkunov, Vladimir M. Blinov, Sergei M. Resenchuk, Alexei V. Totmenin, Viktor N. Krasnykh, Ludmilla V. Olenina, **Oleg I. Serpinsky**, and Lev S. Sandakhchiev

A High-Speed Automated DNA Sequencer
Lloyd M. Smith

Characterization and Modification of DNA Polymerases for Use in DNA Sequencing
Stanley Tabor

Mapping

*Toward Cloning Human Chromosome 19 in Yeast Artificial Chromosomes
Inga P. Arman, Alexander B. Devin, Svetlana P. Legchilina, Irina G. Efimenko, Marina E. Smirnova, and Dina V. Glazkova

A Panel of Mouse-Human Monochromosomal Hybrid Cell Lines, Each Containing a Single Different Tagged Human Chromosome
Arbansjit K. Sandhu, G. Pal Kaur, and **Raghbir S. Athwal**

*Preparation of a Set of Molecular Markers for Human Chromosome 5 Using G+C–Rich and Functional Site-Specific Oligonucleotides
M.L. Filipenko, A.I. Muravlev, E.I. Jantsen, V.V. Smirnova, N.A. Chikaev, V.P. Mishin, and M.A. Ivanovich

An Improved Method for Producing Radiation Hybrids Applied to Human Chromosome 19
Cynthia L. Jackson and Hon Fong L. Mark

Completed Projects

Construction of a Human Genome Library Composed of Multimegabase Acentric Chromosome Fragments
Michael J. Lane, **Peter Hahn**, and **John Hozier**

Reagents for Understanding and Sequencing the Human Genome
J.R. Korenberg, X-N. Chen, S. Mitchell, S. Gerwehr, Z. Sun, D. Noya, R. Hubert, U-J. Kim, H. Shizuya, X. Wu, J. Silva, B. Birren, T.J. Hudson, P. de Jong, E. Lander, and M. Simon

Development of Diallelic Marker Maps Using PCR/OLA
Deborah A. Nickerson and **Pui-Yan Kwok**

Multiplex Mapping of Human cDNAs
William C. Nierman, Donna R. Maglott, and Scott Durkin

Physical Mapping in Preparation for DNA Sequencing
Andreas Gnirke, Regina Lim, Gane Wong, Jun Yu, Roger Bumgarner, and **Maynard Olson**

Construction of a Genetic Map Across Chromosome 21
Elaine A. Ostrander

Integrated Physical Mápping of Human cDNAs
Mihael H. Polymeropoulos

Sequence-Tagged Sites for Human Chromosome 19 cDNAs
Michael J. Siciliano and Anthony V. Carrano

cDNA/STS Map of the Human Genome: Methods Development and Applications Using Brain cDNAs
James M. Sikela, Akbar S. Khan, Arto K. Orpana, Andrea S. Wilcox, Janet A. Hopkins, and Tamara J. Stevens

Physical Structure of Human Chromosome 21
Cassandra L. Smith, Denan Wang, Kaoru Yoshida, Jesus Sainz, Carita Fockler, and Meire Bremer

Physical Mapping of Human Chromosome 16
David F. Callen, Sinoula Apostolou, Elizabeth Baker, Helen Kozman, Sharon A. Lane, Julie Nancarrow, Hilary A. Phillips, Scott A. Whitmore, Norman A. Doggett, John C. Mulley, Robert I. Richards, and **Grant R. Sutherland**

Chromosome Mapping by FISH to Interphase Nuclei
Barbara J. Trask

Flow Karyotyping and Flow Instrumentation Development
Ger van den Engh and **Barbara Trask**

Isolation of Specific Human Telomeric Clones by Homologous Recombination and YAC Rescue
Geoffrey Wahl and Linnea Brody

Informatics

*A Method for Direct Sequencing of Diploid Genomes on Oligonucleotide Arrays: Theoretical Analysis and Computer Modeling
Alexander B. Chetverin

Sampling-Based Methods for the Estimation of DNA Sequence Accuracy
Gary Churchill and **Betty Lazareva**

Computer-Aided Genome Map Assembly with SIGMA (System for Integrated Genome Map Assembly)
Michael J. Cinkosky, Michael A. Bridgers, William M. Barber, Mohamad Ijadi, and James W. Fickett

Informatics for the Sequencing by Hybridization Project
Aleksandar Milosavljevic and **Radomir Crkvenjakov**

Sequencing by Hybridization Algorithms and Computational Tools
Radoje Drmanac, Ivan Labat, and Nick Stavropoulos

HGIR: Information Management for a Growing Map
James W. Fickett, Michael J. Cinkosky, Michael A. Bridgers, Henry T. Brown, Christian Burks, Philip E. Hempfner, Tran N. Lai, Debra Nelson, Robert M. Pecherer, Doug Sorenson, Peichen H. Sgro, Robert D. Sutherland, Charles D. Troup, and Bonnie C. Yantis

Identification of Genes in Anonymous DNA Sequences
Christopher A. Fields and **Carol A. Soderlund**

Algorithms in Support of the Human Genome Project
Dan Gusfield, Jim Knight, Kevin Murphy, Paul Stelling, Lushen Wang, Archie Cobbs, Paul Horton, Richard Karp, and Gene Lawler

BISP: VLSI Solutions to Sequence-Comparison Problems
Tim Hunkapiller, Leroy Hood, Ed Chen, and Michael Waterman

Physical Mapping of DNA Molecules
Richard M. Karp

BIOSCI Electronic Newsgroup Network for the Biological Sciences
David Kristofferson

Multiple Alignment and Homolog Sequence Database Compilation
Hwa A. Lim

Applying Machine Learning Techniques to DNA Sequence Analysis
Jude W. Shavlik, Michiel O. Noordewier, Geoffrey Towell, Mark Craven, Andrew Whitsitt, Kevin Cherkauer, and Lorien Pratt

New Approaches to Recognizing Functional Domains in Biological Sequences
Gary D. Stormo

ELSI

Protecting Genetic Privacy by Regulating the Collection, Analysis, Use, and Storage of DNA and Information Obtained from DNA Analysis
George J. Annas, Leonard H. Glantz, and Patricia A. Roche

"The Secret of Life"
Paula Apsell and **Graham Chedd**

Genome Technology and Its Implications: A Hands-On Workshop for Educators
Diane Baker and **Paula Gregory**

Predicting Future Disease: Issues in the Development, Application, and Use of Tests for Genetic Disorders
Ruth E. Bulger and Jane E. Fullarton

HUGO International Yearbook: Genetics, Ethics, Law, and Society (GELS)
Alex Capron and **Bartha Knoppers**

The Human Genome: Science and the Social Consequences; Interactive Exhibits and Programs on Genetics and the Human Genome
Charles C. Carlson

International Conference Working Group: The Social Costs and Medical Benefits of Human Genetic Information
Betsy Fader

"Medicine at the Crossroads"
George Page and **Stefan Moore**

Pilot Senior Research Fellowship Program: Bioethical Issues in Molecular Genetics
Declan Murphy and **Claudette Cyr Friedman**

Studies of Genetic Discrimination
Marvin Natowicz

DNA Banking and DNA Data Banking: Legal, Ethical, and Public Policy Issues
Philip Reilly

Mechanical Interactive Exhibits on Biotechnology
Elizabeth Sharpe

Impact of Technology Derived from the Human Genome Project on Genetic Testing, Screening, and Counseling: Cultural, Ethical, and Legal Issues
Ralph W. Trottier, Lee A. Crandall, David Phoenix, Mwalimu Imara, and Ray E. Mosley

Social Science Concepts and Studies of Privacy: A Comprehensive Inventory and Analysis for Considering Privacy, Confidentiality, and Access Issues in the Use of Genetic Tests and Applications of Genetic Data
Alan F. Westin

Completed Projects

Human Genetics and Genome Analysis: A Practical Workshop for Public Policymakers and Opinion Leaders
Jan Witkowski, David A. Micklos, and Margaret Henderson

SBIR Phase I

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogenous Public Genome Databases
J. Clarke Anderson

Techniques for Screening Large-Insert Libraries
Saika Aytay

Interactive DNA Sequence Processing for a Microcomputer
Wayne Dettloff and **Holt Anderson**

High-Performance Searching and Pattern Recognition for Human Genome Databases
Douglas J. Eadline

Estimating, Encoding, and Using Uncertainties in Sequence Data
John R. Hartman

Low-Cost Massively Parallel Neurocomputing for Pattern Recognition in Macromolecular Sequences
John R. Hartman

Electrophoretic Separation of DNA Fragments in Ultrathin Planar-Format Linear Polyacrylamide
Michael T. MacDonell and **Darlene B. Roszak**

An Acoustic Plate Mode DNA Biosensor
Douglas J. McAllister

Piezoelectric Biosensor Using Peptide Nucleic Acids for Triplex Capture
Douglas McAllister

Pedigree Software for the Presentation of Human Genome Information for Genetic Education and Counseling
Charles L. Manske

A High-Spatial-Resolution Spectrograph for DNA Sequencing
Cathy D. Newman

Nonradioactive Detection Systems Based on Enzyme-Fragment Complementation
Peter Richterich

Separation Media for DNA Sequencing
David S. Soane and Herbert H. Hooper

SBIR Phase II

Increased Speed in DNA Sequencing by Utilizing LARIS and SIRIS to Localize Multiple Stable Isotope-Labeled Fragments
Heinrich F. Arlinghaus

Rapid, High-Throughput DNA Sequencing Using Confocal Fluorescence Imaging of Capillary Arrays
David L. Barker and **Jay Flatley**

Spatially Defined Oligonucleotide Arrays
Stephen P. A. Fodor

Site-Specific Endonucleases for Human Genome Mapping
George Golumbeski, Kimberly Knoche, Susanne Selman, im Hartnett, Lydia Hung, and Peter Bayne

High-Performance DNA and Protein Sequence Analysis on a Low-Cost Parallel-Processor Array
John R. Hartman and David L. Solomon

Chemiluminescent Multiprimed DNA Sequencing
Chris S. Martin, Corinne E. M. Olesen, and **Irena Bronstein**

Appendix
Narratives from Large, Multidisciplinary Research Projects

Part 1 of this report contains narratives that represent DOE Human Genome Program research in large, multidisciplinary projects. As a convenience to the reader, these narratives are reprinted without graphics in this appendix. Only the contact persons for these organizations are listed in the Index to Principal and Coinvestigators. To obtain more information on research carried out in these projects, see their contact information or visit the Web sites listed with the narratives.

Joint Genome Institute 72
Elbert Branscomb

Lawrence Livermore National Laboratory Human Genome Center 73
Anthony V. Carrano

Los Alamos National Laboratory Center for Human Genome Studies 77
Larry L. Deaven

Lawrence Berkeley National Laboratory Human Genome Center 81
Mohandas Narla

University of Washington Genome Center 85
Maynard Olson

Genome Database 87
Stanley Letovsky and Robert Cottingham

National Center for Genome Resources 91
Peter Schad

Joint Genome Institute
Genome Center Sequencing Efforts Merge

Lawrence Livermore National Laboratory
7000 East Avenue, L-452
Livermore, CA 94551

Elbert Branscomb, JGI Scientific Director
510/422-5681
elbert@alu.llnl.gov* or *elbert@shotgun.llnl.gov

http://www.jgi.doe.gov

In a major restructuring of its Human Genome Program, on October 23, 1996, the DOE Office of Biological and Environmental Research established the Joint Genome Institute (JGI) to integrate work based at its three major human genome centers.

The JGI merger represents a shift toward large-scale sequencing via intensified collaborations for more effective use of the unique expertise and resources at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory. Elbert Branscomb (LLNL) serves as JGI's Scientific Director. Capital equipment has been ordered, and operational support of about $30 million is projected for the 1998 fiscal year.

With easy access to both LBNL and LLNL, a building in Walnut Creek, California, is being modified. Here, starting in late FY 1998, production DNA sequencing will be carried out for JGI. Until that time, large-scale sequencing will continue at LANL, LBNL, and LLNL. Expectations are that within 3 to 4 years the Production Sequencing Facility will house some 200 researchers and technicians working on high-throughput DNA sequencing using state-of-the-art robotics.

Initial plans are to target gene-rich regions of around 1 to 10 megabases for sequencing. Considerations include gene density, gene families (especially clustered families), correlations to model organism results, technical capabilities, and relevance to the DOE mission (e.g., DNA repair, cancer susceptibility, and impact of genotoxins). The JGI program is subject to regular peer review.

Sequence data will be posted daily on the Web; as the information progresses to finished quality, it will be submitted to public databases.

As JGI and other investigators involved in the Human Genome Project are beginning to reveal the DNA sequence of the 3 billion base pairs in a reference human genome, the data already are becoming valuable reagents for explorations of DNA sequence function in the body, sometimes called "functional genomics." Although large-scale sequencing is JGI's major focus, another important goal will be to enrich the sequence data with information about its biological function. One measure of JGI's progress will be its success at working with other DOE laboratories, genome centers, and non-DOE academic and industrial collaborators. In this way, JGI's evolving capabilities can both serve and benefit from the widest array of partners.

Production DNA Sequencing Begun Worldwide

The year 1996 marked a transition to the final and most challenging phase of the U.S. Human Genome Project, as pilot programs aimed at refining large-scale sequencing strategies and resources were funded by DOE and NIH (see Research Highlights, DNA Sequencing, p. 14). Internationally, large-scale human genome sequencing was kicked off in late 1995 when The Wellcome Trust announced a 7-year, $75-million grant to the private Sanger Centre to scale up its sequencing capabilities. French investigators also have announced intentions to begin production sequencing.

Funding agencies worldwide agree that rapid and free release of data is critical. Other issues include sequence accuracy, types of annotation that will be most useful to biologists, and how to sustain the reference sequence.

The international Human Genome Organisation maintains a Web page to provide information on current and future sequencing projects and links to sites of participating groups (http://hugo.gdb.org). The site also links to reports and resources developed at the February 1996 and 1997 Bermuda meetings on large-scale human genome sequencing, which were sponsored by The Wellcome Trust.

Lawrence Livermore National Laboratory Human Genome Center

Human Genome Center
Lawrence Livermore National Laboratory
Biology and Biotechnology Research Program
7000 East Avenue, L-452
Livermore, CA 94551

Anthony V. Carrano, Director
510/422-5698, Fax: /423-3110, *carrano1@llnl.gov*

Linda Ashworth, Assistant to Center Director
510/422-5665, Fax: -2282, *ashworth1@llnl.gov*

http://www-bio.llnl.gov/bbrp/genome/genome.html

The Human Genome Center at Lawrence Livermore National Laboratory (LLNL) was established by DOE in 1991. The center operates as a multidisciplinary team whose broad goal is understanding human genetic material. It brings together chemists, biologists, molecular biologists, physicists, mathematicians, computer scientists, and engineers in an interactive research environment focused on mapping, DNA sequencing, and characterizing the human genome.

Goals and Priorities

In the past 2 years, the center's goals have undergone an exciting evolution. This change is the result of several factors, both intrinsic and extrinsic to the Human Genome Project. They include: (1) successful completion of the center's first-phase goal, namely a high-resolution, sequence-ready map of human chromosome 19; (2) advances in DNA sequencing that allow accelerated scaleup of this operation; and (3) development of a strategic plan for LLNL's Biology and Biotechnology Research Program that will integrate the center's resources and strengths in genomics with programs in structural biology, individual susceptibility, medical biotechnology, and microbial biotechnology.

The primary goal of LLNL's Human Genome Center is to characterize the mammalian genome at optimal resolution and to provide information and material resources to other in-house or collaborative projects that allow exploitation of genomic biology in a synergistic manner. DNA sequence information provides the biological driver for the center's priorities:

- Generation of highly accurate sequence for chromosome 19.
- Generation of highly accurate sequence for genomic regions of high biological interest to the mission of the DOE Office of Biological and Environmental Research (e.g., genes involved in DNA repair, replication, recombination, xenobiotic metabolism, and cell-cycle control).
- Isolation and sequence of the full insert of cDNA clones associated with genomic regions being sequenced.
- Sequence of selected corresponding regions of the mouse genome in parallel with the human.
- Annotation and position of the sequenced clones with physical landmarks such as linkage markers and sequence tagged sites (STSs).
- Generation of mapped chromosome 19 and other genomic clones [cosmids, bacterial artificial chromosomes (BACs), and P1 artificial chromosomes (PACs)] for collaborating groups.
- Sharing of technology with other groups to minimize duplication of effort.
- Support of downstream biology projects, for example, structural biology, functional studies, human variation, transgenics, medical biotechnology, and microbial biotechnology with know-how, technology, and material resources.

Center Organization and Activities

Completion and publication of the metric physical map of human chromosome 19 in 1995 has led to consolidation of many functions associated with physical mapping, with increased emphasis on DNA sequencing. The center is organized into five broad areas of research and support: sequencing, resources, functional genomics, informatics and analytical genomics, and instrumentation. Each area consists of multiple projects, and extensive interaction occurs both within and among projects.

Sequencing

The sequencing group is divided into several subprojects. The core team is responsible for the construction of sequence libraries, sequencing reactions, and data collection for all templates in the random phase of sequencing. The finishing team works with data produced by the core team to produce highly redundant, highly accurate "finish" sequence on targets of interest. Finally, a team of researchers focuses specifically on development, testing, and implementation of new protocols for the entire group, with an emphasis on improving the efficiency and cost basis of the sequencing operation.

LLNL

Resources

The resources group provides mapped clonal resources to the sequencing teams. This group performs physical mapping as needed for the DNA sequencing group by using fingerprinting, restriction mapping, fluorescence in situ hybridization, and other techniques. A small mapping effort is under way to identify, isolate, and characterize BAC clones (from anywhere in the human genome) that relate to susceptibility genes, for example, DNA repair. These clones will be characterized and provided for sequencing and at the same time contribute to understanding the biology of the chromosome, the genome, and susceptibility factors. The mapping team also collaborates with others using the chromosome 19 map as a resource for gene hunting.

Functional Genomics

The functional genomics team is responsible for assembling and characterizing clones for the Integrated Molecular Analysis of Gene Expression (called IMAGE) Consortium and cDNA sequencing, as well as for work on gene expression and comparative mouse genomics. The effort emphasizes genes involved in DNA repair and links strongly to LLNL's gene-expression and structural biology efforts. In addition, this team is working closely with Oak Ridge National Laboratory (ORNL) to develop a comparative map and the sequence data for mouse regions syntenic to human chromosome 19.

Informatics and Analytical Genomics

The informatics and analytical genomics group provides computer science support to biologists. The sequencing informatics team works directly with the DNA sequencing group to facilitate and automate sample handing, data acquisition and storage, and DNA sequence analysis and annotation. The analytical genomics team provides statistical and advanced algorithmic expertise. Tasks include development of model-based methods for data capture, signal processing, and feature extraction for DNA sequence and fingerprinting data and analysis of the effectiveness of newly proposed methods for sequencing and mapping.

Instrumentation

The instrumentation group also has multiple components. Group members provide expertise in instrumentation and automation in high-throughput electrophoresis, preparation of high-density replicate DNA and colony filters, fluorescence labeling technologies, and automated sample handling for DNA sequencing. To facilitate seamless integration of new technologies into production use, this group is coupled tightly to the biologist user groups and the informatics group.

Collaborations

The center interacts extensively with other efforts within the LLNL Biology and Biotechnology Research Program and with other programs at LLNL, the academic community, other research institutes, and industry. More than 250 collaborations range from simple probe and clone sharing to detailed gene family studies. The following list reflects some major collaborations.

- Integration of the genetic map of human chromosome 19 with corresponding mouse chromosomes (ORNL).
- Miniaturized polymerase chain reaction instrumentation (LLNL).
- Sequencing of IMAGE Consortium cDNA clones (Washington University, St. Louis).
- Mapping and sequencing of a gene associated with Finnish congenital nephrotic syndrome (University of Oulu, Finland).

Accomplishments

The LLNL Human Genome Center has excelled in several areas, including comparative genomic sequencing of DNA repair genes in human and rodent species, construction of a metric physical map of human chromosome 19, and development and application of new biochemical and mathematical approaches for constructing ordered clone maps. These and other major accomplishments are highlighted below.

- Completion of highly accurate sequencing totaling 1.6 million bases of DNA, including regions spanning human DNA repair genes, the candidate region for a congenital kidney disease gene, and other regions of biological interest on chromosome 19.
- Completion of comparative sequence analysis of 107,500 bases of genomic DNA encompassing the human DNA repair gene *ERCC*2 and the corresponding regions in mouse and hamster. In addition to *ERCC*2, analysis revealed the presence of two previously undescribed genes in all three species. One of these genes is a new member of the kinesin motor protein family. These proteins play a wide variety of roles in the cell, including movement of chromosomes before cell division.
- Complete sequencing of human genomic regions containing two additional DNA repair genes. One of these, *XRCC*3, maps to human chromosome 14 and encodes a protein that may be required for chromosome stability. Analysis of the genomic sequence identified another kinesin motor protein gene physi-

cally linked to *XRCC*3. The second human repair gene, *HHR*23A, maps to 19p13.2. Sequence analysis of 110,000 bases containing *HHR*23A identified six other genes, five of which are new genes with similarity to proteins from mouse, human, yeast, and *Caenorhabditis elegans*.

- Complete sequencing of full-length cDNAs for three new DNA repair genes (*XRCC*2, *XRCC*3, and *XRCC*9) in collaboration with the LLNL DNA repair group.
- Generation of a metric physical map of chromosome 19 spanning at least 95% of the chromosome. This unique map incorporates a metric scale to estimate the distance between genes or other markers of interest to the genetics community.
- Assembly of nearly 45 million bases of *Eco*R I restriction-mapped cosmid contigs for human chromosome 19 using a combination of fingerprinting and cosmid walking. Small gaps in cosmid continuity have been spanned by BAC, PAC, and P1 clones, which are then integrated into the restriction maps. The high depth of coverage of these maps (average redundancy, 4.3-fold) permits selection of a minimum overlapping set of clones for DNA sequencing.
- Placement of more than 400 genes, genetic markers, and other loci on the chromosome 19 cosmid map. Also, 165 new STSs associated with premapped cosmid contigs were generated and added to the physical map.
- Collaborations to identify the gene (*COMP*) responsible for two allelic genetic diseases, pseudoachondroplasia and multiple epiphyseal dysplasia, and the identification of specific mutations causing each condition.
- Through sequence analysis of the 2A subfamily of the human cytochrome P450 enzymes, identification of a new variant that exists in 10% to 20% of individuals and results in reduced ability to metabolize nicotine and the antiblood-clotting drug Coumadin.
- Location of a zinc finger gene that encodes a transcription factor regulating blood-cell development adjacent to telomere repeat sequences, possibly the gene nearest one end of chromosome 19.
- Completion of the genomic and cDNA sequence of the gene for the human Rieske Fe-S protein involved in mitochondrial respiration.
- Expansion of the mouse-human comparative genomics collaboration with ORNL to include study of new groups of clustered transcription factors found on human chromosome 19q and as syntenic homologs on mouse chromosome 7.
- Numerous collaborations (in particular, with Washington University and Merck) continuing to expand the LLNL-based IMAGE Consortium, an effort to characterize the transcribed human genome. The IMAGE clone collection is now the largest public collection of sequenced cDNA clones, with more than 500,000 arrayed clones, 500,000 sequences in public databases, and 10,000 mapped cDNAs.
- Development and deployment of a comprehensive system to handle sample tracking needs of production DNA sequencing. The system combines databases and graphical interfaces running on both Mac and Sun platforms and scales easily to handle large-scale production sequencing.
- Expansion of the LLNL genome center's World Wide Web site to include tables that link to each gene being sequenced, to the quality scores and assembled bases collected each night during the sequencing process, and to the submitted GenBank sequence when a clone is completed. [*http://bbrp.llnl.gov/test-bin/projqcsummary*]
- Implementation of a new database to support sequencing and mapping work on multiple chromosomes and species. Web-based automated tools were developed to facilitate construction of this database, the loading of over 100 million bytes of chromosome 19 data from the existing LLNL database, and automated generation of Web-based input interfaces.
- Significant enhancement of the LLNL Genome Graphical Database Browser software to display and link information obtained at a subcosmid resolution from both restriction map hybridization and sequence feature data. Features, such as genes linked to diseases, allow tracking to fragments as small as 500 base pairs of DNA.
- Development of advanced microfabrication technologies to produce electrophoresis microchannels in large glass substrates for use in DNA sequencing.
- Installation of a new filter-spotting robot that routinely produces 6 × 6 × 384 filters. A 16 × 16 × 384 pattern has been achieved.
- Upgrade of the Lawrence Berkeley National Laboratory colony picker using a second computer so that imaging and picking can occur simultaneously.

Future Plans

Genomic sequencing currently is the dominant function of Livermore's Human Genome Center. The physical mapping effort will ensure an ample supply of sequence-ready clones. For sequencing targets on chromosome 19, this includes ensuring that the most stable clones (cosmids, BACs, and PACs) are available for sequencing and that regions with such known physical landmarks as STSs and expressed sequenced tags (ESTs) are annotated to facilitate sequence assembly and analysis. The following targets are emphasized for DNA sequencing:

- Regions of high gene density, including regions containing gene families.
- Chromosome 19, of which at least 42 million bases are sequence ready.
- Selected BAC and PAC clones representing regions of about 0.2 million to 1 million bases throughout the human genome; clones would be selected based on such high-priority biological targets as genes involved in DNA repair, replication, recombination, xenobiotic metabolism, cell-cycle checkpoints, or other specific targets of interest.
- Selected BAC and PAC clones from mouse regions syntenic with the genes indicated above.
- Full-insert cDNAs corresponding to the genomic DNA being sequenced.

The informatics team is continuing to deploy broader-based supporting databases for both mapping and sequencing. Where appropriate, Web- and Java-based tools are being developed to enable biologists to interact with data. Recent reorganization within this group enables better direct support to the sequencing group, including evaluating and interfacing sequence-assembly algorithms and analysis tools, data and process tracking, and other informatics functions that will streamline the sequencing process.

The instrumentation effort has three major thrusts: (1) continued development or implementation of laboratory automation to support high-throughput sequencing; (2) development of the next-generation DNA sequencer; and (3) development of robotics to support high-density BAC clone screening. The last two goals warrant further explanation.

The new DNA sequencer being developed under a grant from the National Institutes of Health, with minor support through the DOE genome center, is designed to run 384 lanes simultaneously with a low-viscosity sieving medium. The entire system would be loaded automatically, run, and set up for the next run at 3-hour intervals. If successful, it should provide a 20- to 40-fold increase in throughput over existing machines.

An LLNL-designed high-precision spotting robot, which should allow a density of 98,304 spots in 96 cm^2, is now operating. The goal of this effort is to create high-density filters representing a 10× BAC coverage of both human and mouse genomes (30,000 clones = 1× coverage). Thus each filter would provide ~3× coverage, and eight such filters would provide the desired coverage for both genomes. The filters would be hybridized with amplicons from individual or region-specific cDNAs and ESTs; given the density of the BAC libraries, clones that hybridize should represent a binned set of BACs for a region of interest. These BACs could be the initial substrate for a BAC sequencing strategy. Performing hybridizations in parallel in mouse and human DNA facilitates the development of the mouse map (with ORNL involvement), and sequencing BACs from both species identifies evolutionarily conserved and, perhaps, regulatory regions.

Information generated by sequencing human and mouse DNA in parallel is expected to expand LLNL efforts in functional genomics. Comparative sequence data will be used to develop a high-resolution synteny map of conserved mouse-human domains and incorporate automated northern expression analysis of newly identified genes. Long range, the center hopes to take advantage of a variety of forms of expression analysis, including site-directed mutation analysis in the mouse.

Summary

The Livermore Human Genome Center has undergone a dramatic shift in emphasis toward commitment to large-scale, high-accuracy sequencing of chromosome 19, other chromosomes, and targeted genomic regions in the human and mouse. The center also is committed to exploiting sequence information for functional genomics studies and for other programs, both in house and collaboratively.

Los Alamos National Laboratory Center for Human Genome Studies

Center for Human Genome Studies
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545

Robert K. Moyzis, Director, 1989–97*

Larry L. Deaven, Acting Director
505/667-3912, Fax: -2891
ldeaven@telomere.lanl.gov

Lynn Clark, Technical Coordinator
505/667-9376, Fax: -2891
clark@telomere.lanl.gov

http://www-ls.lanl.gov/masterhgp.html

Biological research was initiated at Los Alamos National Laboratory (LANL) in the 1940s, when the laboratory began to investigate the physiological and genetic consequences of radiation exposure. Eventual establishment of the national genetic sequence databank called GenBank, the National Flow Cytometry Resource, numerous related individual research projects, and fulfillment of a key role in the National Laboratory Gene Library Project all contributed to LANL's selection as the site for the Center for Human Genome Studies in 1988.

Center Organization and Activities

The LANL genome center is organized into four broad areas of research and support: Physical Mapping, DNA Sequencing, Technology Development, and Biological Interfaces. Each area consists of a variety of projects, and work is distributed among five LANL Divisions (Life Sciences; Theoretical; Computing, Information, and Communications; Chemical Science and Technology; and Engineering Sciences and Applications). Extensive interdisciplinary interactions are encouraged.

Physical Mapping

The construction of chromosome- and region-specific cosmid, bacterial artificial chromosome (BAC), and yeast artificial chromosome (YAC) recombinant DNA libraries is a primary focus of physical mapping activities at LANL. Specific work includes the construction of high-resolution maps of human chromosomes 5 and 16 and associated informatics and gene discovery tasks.

Accomplishments

- Completion of an integrated physical map of human chromosome 16 consisting of both a low-resolution YAC contig map and a high-resolution cosmid contig map. With sequence tagged site (STS) markers provided on average every 125,000 bases, the YAC-STS map provides almost-complete coverage of the chromosome's euchromatic arms. All available loci continue to be incorporated into the map.
- Construction of a low-resolution STS map of human chromosome 5 consisting of 517 STS markers regionally assigned by somatic-cell hybrid approaches. Around 95% mega-YAC–STS coverage (50 million bases) of 5p has been achieved. Additionally, about 40 million bases of 5q mega-YAC–STS coverage have been obtained collaboratively.
- Refinement of BAC cloning procedures for future production of chromosome-specific libraries. Successful partial digestion and cloning of microgram quantities of chromosomal DNA embedded in agarose plugs. Efforts continue to increase the average insert size to about 100,000 bases.

DNA Sequencing

DNA sequencing at the LANL center focuses on low-pass sample sequencing (SASE) of large genomic regions. SASE data is deposited in publicly available databases to allow for wide distribution. Finished sequencing is prioritized from initial SASE analysis and pursued by parallel primer walking. Informatics development includes data tracking, gene-discovery integration with the Sequence Comparison ANalysis (SCAN) program, and functional genomics interaction.

Accomplishments

- SASE sequencing of 1.5 million bases from the p13 region of human chromosome 16.
- Discovery of more than 100 genes in SASE sequences.
- Generation of finished sequence for a 240,000-base telomeric region of human chromosome 7q. From initial sequences generated by SASE, oligonucleotides were synthesized and used for primer walking directly from cosmids comprising the contig map. Complete sequencing was performed to determine what genes, if any, are near the 7q terminus. This intriguing region lacks significant blocks of subtelomeric repeat DNA typically present near eukaryotic telomeres.

---

*Now at University of California, Irvine

- Complete single-pass sequencing of 2018 exon clones generated from LANL's flow-sorted human chromosome 16 cosmid library. About 950 discrete sequences were identified by sequence analysis. Nearly 800 appear to represent expressed sequences from chromosome 16.
- Development of Sequence Viewer to display ABI sequences with trace data on any computer having an Internet connection and a Netscape World Wide Web browser.
- Sequencing and analysis of a novel pericentromeric duplication of a gene-rich cluster between 16p11.1 and Xq28 (in collaboration with Baylor College of Medicine).

Technology Development

Technology development encompasses a variety of activities, both short and long term, including novel vectors for library construction and physical mapping; automation and robotics tools for physical mapping and sequencing; novel approaches to DNA sequencing involving single-molecule detection; and novel approaches to informatics tools for gene identification.

Accomplishments

- Development of SCAN program for large-scale sequence analysis and annotation, including a translator converting SCAN data to GIO format for submission to Genome Sequence DataBase.
- Application of flow-cytometric approach to DNA sizing of P1 artificial chromosome (PAC) clones. Less than one picogram of linear or supercoiled DNA is analyzed in under 3 minutes. Sizing range has been extended down to 287 base pairs. Efforts continue to extend the upper limit beyond 167,000 bases.
- Characterization of the detection of single, fluorescently tagged nucleotides cleaved from multiple DNA fragments suspended in the flow stream of a flow cytometer. The cleavage rate for Exo III at 37°C was measured to be about 5 base pairs per second per M13 DNA fragment. To achieve a single-color sequencing demonstration, either the background burst rate (currently about 5 bursts per second) must be reduced or the exonuclease cleavage rate must be increased significantly. Techniques to achieve both are being explored.
- Construction of a simple and compact apparatus, based on a diode-pumped Nd:YAG laser, for routine DNA fragment sizing.
- Development of a new approach to detect coding sequences in DNA. This complete spectral analysis of coding and noncoding sequences is as sensitive in its first implementations as the best existing techniques.
- Use of phylogenetic relationships to generate new profiles of amino acid usage in conserved domains. The profiles are particularly useful for classification of distantly related sequences.

Biological Interfaces

The Biological Interfaces effort targets genes and chromosome regions associated with DNA damage and repair, mitotic stability, and chromosome structure and function as primary subjects for physical mapping and sequencing. Specific disease-associated genes on human chromosome 5 (e.g., Cri-du-Chat syndrome) and on 16 (e.g., Batten's disease and Fanconi anemia) are the subjects of collaborative biological projects.

Accomplishments

- Identification of two human 7q exons having 99% homology to the cDNA of a known human gene, vasoactive intestinal peptide receptor 2A. Preliminary data suggests that the *VIPR*2A gene is expressed.
- Identification of numerous expressed sequence tags (ESTs) localized to the 7q region. Since three of the ESTs contain at least two regions with high confidence of homology (~90%), genes in addition to *VIPR*2A may exist in the terminal region of 7q.
- Generation of high-resolution cosmid coverage on human chromosome 5p for the larynx and critical regions identified with Cri-du-Chat syndrome, the most common human terminal-deletion syndrome (in collaboration with Thomas Jefferson University).
- Refinement of the Wolf-Hirschhorn syndrome (WHS) critical region on human chromosome 4p. Using the SCAN program to identify genes likely to contribute to WHS, the project serves as a model for defining the interaction between genomic sequencing and clinical research.
- Collaborative construction of contigs for human chromosome 16, including 1.05 million bases in cosmids through the familial Mediterranean fever (FMF) gene region (with members of the FMF Consortium) and 700,000 bases in P1 clones encompassing the polycystic kidney disease gene (with Integrated Genetics, Inc.).
- Collaborative identification and determination of the complete genomic structure of the Batten's disease gene (with members of the BDG Consortium), the gamma subunit of the human amiloride-sensitive epithelial channel (Liddle's syndrome, with University of Iowa), and the polycystic kidney disease gene (with Integrated Genetics).

- Participation in an international collaborative research consortium that successfully identified the gene responsible for Fanconi anemia type A.

Patents, Licenses, and CRADAs

- Rhett L. Affleck, James N. Demas, Peter M. Goodwin, Jay A. Schecker, Ming Wu, and Richard A. Keller, "Reduction of Diffusional Defocusing in Hydrodynamically Focused Flows by Complexing with a High Molecular Weight Adduct," United States Patent, filed December 1996.
- R.L. Affleck, W.P. Ambrose, J.D. Demas, P.M. Goodwin, M.E. Johnson, R.A. Keller, J.T. Petty, J.A. Schecker, and M. Wu, "Photobleaching to Reduce or Eliminate Luminescent Impurities for Ultrasensitive Luminescence Analysis," United States Patent, S-87, 208, accepted September 1997.
- J.H. Jett, M.L. Hammond, R.A. Keller, B.L. Marrone, and J.C. Martin, "DNA Fragment Sizing and Sorting by Laser-Induced Fluorescence," United States Patent, S.N. 75,001, allowed May 1996.
- James H. Jett, "Method for Rapid Base Sequencing in DNA and RNA with Three Base Labeling," in preparation.
- Development license and exclusive license to LANL's DNA sizing patent obtained by Molecular Technology, Inc., for commercialization of single-molecule detection capability to DNA sizing.

Future Plans

LANL has joined a collaboration with California Institute of Technology and The Institute for Genomic Research to construct a BAC map of the p arm of human chromosome 16 and to complete the sequence of a 20-million–base region of this map.

In its evolving role as part of the new DOE Joint Genome Institute, LANL will continue scaleup activities focused on high-throughput DNA sequencing. Initial targets include genes and DNA regions associated with chromosome structure and function, syntenic break-points, and relevant disease-gene loci.

A joint DNA sequencing center was established recently by LANL at the University of New Mexico. This facility is responsible for determining the DNA sequence of clones constructed at LANL, then returning the data to LANL for analysis and archiving.

Lawrence Berkeley National Laboratory Human Genome Center

Human Genome Center
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720

Michael Palazzolo,* Director, 1996–97

Contact: Mohandas Narla
510/486-7029, Fax: -6746
mohandas_narla@macmail.lbl.gov

Joyce Pfeiffer, Administrative Assistant

http://www-hgc.lbl.gov/GenomeHome.html

Since 1937 the Ernest Orlando Lawrence Berkeley National Laboratory (LBNL) has been a major contributor to knowledge about human health effects resulting from energy production and use. That was the year John Lawrence went to Berkeley to use his brother Ernest's cyclotrons to launch the application of radioactive isotopes in biological and medical research. Fifty years later, Berkeley Lab's Human Genome Center was established.

Now, after another decade, an expansion of biological research relevant to Human Genome Project goals is being carried out within the Life Sciences Division, with support from the Information and Computing Sciences and Engineering divisions. Individuals in these research projects are making important new contributions to the key fields of molecular, cellular, and structural biology; physical chemistry; data management; and scientific instrumentation. Additionally, industry involvement in this growing venture is stimulated by Berkeley Lab's location in the San Francisco Bay area, home to the largest congregation of biotechnology research facilities in the world.

In July 1997 the Berkeley genome center became part of the Joint Genome Institute.

Sequencing

Large-scale genomic sequencing has been a central, ongoing activity at Berkeley Lab since 1991. It has been funded jointly by DOE (for human genome production sequencing and technology development) and the NIH National Human Genome Research Institute [for sequencing the *Drosophila melanogaster* model system, which is carried out in partnership with the University of California, Berkeley (UCB)]. The human genome sequencing area at Berkeley Lab consists of five groups: Bioinstrumentation, Automation, Informatics, Biology, and Development. Complementing these activities is a group in Life Sciences Division devoted to functional genomics, including the transgenics program.

The directed DNA sequencing strategy at Berkeley Lab was designed and implemented to increase the efficiency of genomic sequencing. A key element of the directed approach is maintaining information about the relative positions of potential sequencing templates throughout the entire sequencing process. Thus, intelligent choices can be made about which templates to sequence, and the number of selected templates can be kept to a minimum. More important, knowledge of the interrelationship of sequencing runs guides the assembly process, making it more resistant to difficulties imposed by repeated sequences. As of July 3, 1997, Berkeley Lab had generated 4.4 megabases of human sequence and, in collaboration with UCB, had tallied 7.6 megabases of *Drosophila* sequence.

Instrumentation and Automation

The instrumentation and automation program encompasses the design and fabrication of custom apparatus to facilitate experiments, the programming of laboratory robots to automate repetitive procedures, and the development of (1) improved hardware to extend the applicability range of existing commercial robots and (2) an integrated operating system to control and monitor experiments. Although some discrete instrumentation modules used in the integrated protocols are obtained commercially, LBNL designs its own custom instruments when existing capabilities are inadequate. The instrumentation modules are then integrated into a large system to facilitate large-scale production sequencing. In addition, a significant effort is devoted to improving fluorescence-assay methods, including DNA sequence analysis and mass spectrometry for molecular sizing.

Recent advances in the instrumentation group include DNA Prep machine and Prep Track. These instruments are designed to automate completely the highly repetitive and labor-intensive DNA-preparation procedure to provide higher daily throughput and DNA of consistent quality for sequencing (see Web pages: *http://hgighub.lbl.gov/esd/DNAPrep/TitlePage.html* and *http://hgighub.lbl.gov/esd/repTrackWebpage/preptrack.htm*).

Berkeley Lab's near-term needs are for 960 samples per day of DNA extracted from overnight bacteria growths. The DNA protocol is a modified boil prep prepared in a 96-well format. Overnight bacteria growths are lysed, and samples are separated from cell debris by centrifugation. The DNA is recovered by ethanol precipitation.

---

*Now at Amgen, Inc.

Informatics

The informatics group is focused on hardware and software support and system administration, software development for end sequencing, transposon mapping and sequence template selection, data-flow automation, gene finding, and sequence analysis. Data-flow automation is the main emphasis. Six key steps have been identified in this process, and software is being written and tested to automate all six. The first step involves controlling gel quality, trimming vector sequence, and storing the sequences in a database. A program module called Move-Track-Trim, which is now used in production, was written to handle these steps. The second through fourth steps in this process involve assembling, editing, and reconstructing P1 clones of 80,000 base pairs from 400-base traces. The fifth step is sequence annotation, and the sixth is data submission.

Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, and gene signals such as promoters. LBNL is developing a workbench for automatic sequence annotation and annotation viewing and editing. The goal is to run a series of sequence-analysis tools and display the results to compare the various predictions. Researchers then will be able to examine all the annotations (for example, genes predicted by various gene-finding methods) and select the ones that look best.

Nomi Harris developed Genotator, an annotation workbench consisting of a stand-alone annotation browser and several sequence-analysis functions. The back end runs several gene finders, homology searches (using BLAST), and signal searches and saves the results in ".ace" format. Genotator thus automates the tedious process of operating a dozen different sequence-analysis programs with many different input and output formats. Genotator can function via command-line arguments or with the graphical user interface (*http://www-hgc.lbl.gov/inf/annotation.html*).

Progress to Date

Chromosome 5

Over the last year, the center has focused its production genomic sequencing on the distal 40 megabases of the human chromosome 5 long arm. This region was chosen because it contains a cluster of growth factor and receptor genes and is likely to yield new and functionally related genes through long-range sequence analysis. Results to date include:

- 40-megabase nonchimeric map containing 82 yeast artificial chromosomes (YACs) in the chromosome 5 distal long arm.
- 20-megabase contig map in the region of 5q23-q33 that contains 198 P1s, 60 P1 artificial chromosomes, and 495 bacterial artificial chromosomes (BACs) linked by 563 sequenced tagged sites (STSs) to form contigs.
- 20-megabase bins containing 370 BACs in 74 bins in the region of 5q33-q35.

Chromosome 21

An early project in the study of Down syndrome (DS), which is characterized by chromosome 21 trisomy, constructed a high-resolution clone map in the chromosome 21 DS region to be used as a pilot study in generating a contiguous gene map for all of chromosome 21. This project has integrated P1 mapping efforts with transgenic studies in the Life Sciences Division. P1 maps provide a suitable form of genomic DNA for isolating and mapping cDNA.

- 186 clones isolated in the major DS region of chromosome 21 comprising about 3 megabases of genomic DNA extending from D21S17 to ETS2. Through cross-hybridization, overlapping P1s were identified, as well as gaps between two P1 contigs, and transgenic mice were created from P1 clones in the DS region for use in phenotypic studies.

Transgenic Mice

One of the approaches for determining the biological function of newly identified genes uses YAC transgenic mice. Human sequence harbored by YACs in transgenic mice has been shown to be correctly regulated both temporally and spatially. A set of nonchimeric overlapping YACs identified from the 5q31 region has been used to create transgenic mice. This set of transgenic mice, which together harbor 1.5 megabases of human sequence, will be used to assess the expression pattern and potential function of putative genes discovered in the 5q31 region. Additional mapping and sequencing are under way in a region of human chromosome 20 amplified in certain breast tumor cell lines.

Resource for Molecular Cytogenetics

Divining landmarks for human disease amid the enormous plain of the human genetic map is the mission of an ambitious partnership among the Berkeley Lab; University of California, San Francisco; and a diagnostics company. The collaborative Resource for Molecular Cytogenetics is charting a course toward important sites of biological interest on the 23 pairs of human chromosomes (*http://rmc-www.lbl.gov*).

The Resource employs the many tools of molecular cytogenetics. The most basic of these tools, and the cornerstone of the Resource's portfolio of proprietary technology, is a method generally known as "chromosome painting," which uses a technique referred to as fluorescence in situ hybridization or FISH. This technology was invented by LBNL Resource leaders Joe Gray and Dan Pinkel.

A technology to emerge recently from the Resource is known as "Quantitative DNA Fiber Mapping (QDFM)." High-resolution human genome maps in a form suitable for DNA sequencing traditionally have been constructed by various methods of fingerprinting, hybridization, and identification of overlapping STSs. However, these techniques do not readily yield information about sequence orientation, the extent of overlap of these elements, or the size of gaps in the map. Ulli Weier of the Resource developed the QDFM method of physical map assembly that enables the mapping of cloned DNA directly onto linear, fully extended DNA molecules. QDFM allows unambiguous assembly of critical elements leading to high-resolution physical maps. This task now can be accomplished in less than 2 days, as compared with weeks by conventional methods. QDFM also enables detection and characterization of gaps in existing physical maps—a crucial step toward completing a definitive human genome map.

University of Washington Genome Center
Department of Medicine
Box 352145
Seattle, WA 98195

Maynard Olson, Director
206/685-7366, Fax: -7344
mvo@u.washington.edu

http://www.genome.washington.edu

The Human Genome Project soon will need to increase rapidly the scale at which human DNA is analyzed. The ultimate goal is to determine the order of the 3 billion bases that encode all heritable information. During the 20 years since effective methods were introduced to carry out DNA sequencing by biochemical analysis of recombinant-DNA molecules, these techniques have improved dramatically. In the late 1970s, segments of DNA spanning a few thousand bases challenged the capacity of world-class sequencing laboratories. Now, a few million base pairs per year represent state-of-the-art output for a single sequencing center.

However, the Human Genome Project is directed toward completing the human sequence in 5 to 10 years, so the data must be acquired with technology available now. This goal, while clearly feasible, poses substantial organizational and technical challenges. Organizationally, genome centers must begin building data-production units capable of sustained, cost-effective operation. Technically, many incremental refinements of current technology must be introduced, particularly those that remove impediments to increasing the scale of DNA sequencing. The University of Washington (UW) Genome Center is active in both areas.

Production Sequencing

Both to gain experience in the production of high-quality, low-cost DNA sequence and to generate data of immediate biological interest, the center is sequencing several regions of human and mouse DNA at a current throughput of 2 million bases per year. This "production sequencing" has three major targets: the human leukocyte antigen (HLA) locus on human chromosome 6, the mouse locus encoding the alpha subunit of T-cell receptors, and an "anonymous" region of human chromosome 7.

The HLA locus encodes genes that must be closely matched between organ donors and organ recipients. This sequence data is expected to lead to long-term improvements in the ability to achieve good matches between unrelated organ donors and recipients.

The mouse locus that encodes components of the T-cell–receptor family is of interest for several reasons. The locus specifies a set of proteins that play a critical role in cell-mediated immune responses. It provides sequence data that will help in the design of new experimental approaches to the study of immunity in mice—one of the most important experimental animals for immunological research. In addition, the locus will provide one of the first large blocks of DNA sequence for which both human and mouse versions are known.

Human-mouse sequence comparisons provide a powerful means of identifying the most important biological features of DNA sequence because these features are often highly conserved, even between such biologically different organisms as human and mouse. Finally, sequencing an "anonymous" region of human chromosome 7, a region about which little was known previously, provides experience in carrying out large-scale sequencing under the conditions that will prevail throughout most of the Human Genome Project.

Technology for Large-Scale Sequencing

In addition to these pilot projects, the UW Genome Center is developing incremental improvements in current sequencing technology. A particular focus is on enhanced computer software to process raw data acquired with automated laboratory instruments that are used in DNA mapping and sequencing. Advanced instrumentation is commercially available for determining DNA sequence via the "four-color–fluorescence method," and this instrumentation is expected to carry the main experimental load of the Human Genome Project. Raw data produced by these instruments, however, require extensive processing before they are ready for biological analysis.

Large-scale sequencing involves a "divide-and-conquer" strategy in which the huge DNA molecules present in human cells are broken into smaller pieces that can be propagated by recombinant-DNA methods. Individual analyses ultimately are carried out on segments of less than 1000 bases. Many such analyses, each of which still contains numerous errors, must be melded together to obtain finished sequence. During the melding, errors in individual analyses must be recognized and corrected. In typical large-scale sequencing projects, the results of thousands of analyses are melded to produce highly accurate sequence (less than one error in 10,000 bases) that is continuous in blocks of 100,000 or more bases. The UW Genome Center is playing a major role in developing software that allows this process to be carried out automatically with little need for expert intervention. Software developed in the UW center is used in more than 50 sequencing laboratories around the world, including most of the large-scale sequencing centers producing data for the Human Genome Project.

High-Resolution Physical Mapping

The UW Genome Center also is developing improved software that addresses a higher-level problem in large-scale sequencing. The starting point for large-scale sequencing typically is a recombinant-DNA molecule that allows propagation of a particular human genomic segment spanning 50,000 to 200,000 bases. Much effort during the last decade has gone into the physical mapping of such molecules, a process that allows huge regions of chromosomes to be defined in terms of sets of overlapping recombinant-DNA molecules whose precise positions along the chromosome are known. However, the precision required for knowing relationships of recombinant-DNA molecules derived from neighboring chromosomal portions increases as the Human Genome Project shifts its emphasis from mapping to sequencing.

High-resolution maps both guide the orderly sequencing of chromosomes and play a critical role in quality control. Only by mapping recombinant-DNA molecules at high resolution can subtle defects in particular molecules be recognized. Such defective human DNA sources, which are not faithful replicas of the human genome, must be weeded out before sequencing can begin. The UW Genome Center has a major program in high-resolution physical mapping which, like the work on sequencing itself, uses advanced computing tools. The center is producing maps of regions targeted for sequencing on a just-in-time basis. These highly detailed maps are proving extremely valuable in facilitating the production of high-quality sequence.

Ultimate Goal

Although many challenges currently posed by the Human Genome Project are highly technical, the ultimate goal is biological. The project will deliver immense amounts of high-quality, continuous DNA sequence into publicly accessible databases. These data will be annotated so that biologists who use them will know the most likely positions of genes and have convenient access to the best available clues about the probable function of these genes. The better the technical solutions to current challenges, the better the center will be able to serve future users of the human genome sequence.

Genome Database
Johns Hopkins University
2024 E. Monument Street
Baltimore, MD 21205-2236

David Kingsbury, Director, 1993–97*

Stanley Letovsky, Informatics Director
letovsky@gdb.org

Robert Cottingham, Operations Director
bc@gdb.org

Telephone for both: 410/955-9705
Fax for both: 410/614-0434

http://www.gdb.org

The release of Version 6 of the Genome Database (GDB) in January 1996 signaled a major change for both the scientific community and GDB staff. GDB 6.0 introduced a number of significant improvements over previous versions of GDB, most notably a revised data representation for genes and genomic maps and a new curatorial model for the database. These new features, along with a remodeled database structure and new schema and user interface, provide a resource with the potential to integrate all scientific information currently available on human genomics. GDB rapidly is becoming the international biomedical research community's central source for information about genomic structure, content, diversity, and evolution.

A New Data Model

Inherent in the underlying organization of information in GDB is an improved model for genes, maps, and other classes of data. In particular, genomic segments (any named region of the genome) and maps are being expanded regularly. New segment types have been added to support the integration of mapping and sequencing data (for example, gene elements and repeats) and the construction of comparative maps (syntenic regions). New map types include comparative maps for representing conserved syntenies between species and comprehensive maps that combine data from all the various submitted maps within GDB to provide a single integrated view of the genome. Experimental observations such as order, size, distance, and chimerism are also available.

Through the World Wide Web, GDB links its stored data with many other biological resources on the Internet. GDB's External Link category is a growing collection of cross-references established between GDB entities and related information in other databases. By providing a place for these cross-references, GDB can serve as a central point of inquiry into technical data regarding human genomics.

Direct Community Data Submission and Curation

Two methods for data submission are in use. For individuals submitting small amounts of data, interactive editing of the database through the Web became available in April 1996, and the process has undergone several simplifications since that time. This continues to be an area of development for GDB because all editing must take place at the Baltimore site, and Internet connections from outside North America may be too slow for interactive editing to be practical. Until these difficulties are resolved, GDB encourages scientists with limited connectivity to Baltimore to submit their data via more traditional means (e-mail, fax, mail, phone) or to prepare electronic submissions for entry by the data group on site.

For centers submitting large quantities of data, GDB developed an electronic data submission (EDS) tool, which provides the means to specify login password validation and commands for inserting and updating data in GDB. The EDS syntax includes a mechanism for relating a center's local naming conventions to GDB objects. Data submitted to GDB may be stored privately for up to 6 months before it automatically becomes public. The database is programmed to enforce this Human Genome Project policy. Detailed specifications of GDB's EDS syntax and other submission instructions are available (EDS prototype, *http://www.gdb.org/eds*).

Since the EDS system was implemented, GDB has put forth an aggressive effort to increase the amount of data stored in the database. Consequently, the database has grown tremendously. During 1996 it grew from 1.8 to 6.7 gigabytes.

To provide accountability regarding data quality, the shift to community curation introduced the idea that individuals and laboratories own the data they submit to GDB and that other researchers cannot modify it. However, others should be able to add information and comments, so an additional feature is the commu-nity's ability to conduct electronic online public discussions by annotating the database submissions of fellow researchers. GDB is the first database of its kind to offer this feature, and the number of third-party annotations is increasing in the form of editorial commentary, links to literature citations, and links to other databases external to GDB. These links are an important part of the curatorial process because they make other data collections available to GDB users in an appropriate context.

---

*Now at Chiron Pharmaceuticals, Emeryville, California

Improved Map Representation and Querying

Accompanying the release of GDB 6.0, the program Mapview creates graphical displays of maps. Mapview was developed at GDB to display a number of map types (cytogenetic, radiation hybrid, contig, and linkage) using common graphical conventions found in the literature. Mapview is designed to stand alone or to be used in conjunction with a Web browser such as Netscape, thereby creating an interactive graphical display system. When used with Netscape, Mapview allows the user to retrieve details about any displayed map object.

Maps are accessed through the query form for genomic segment and its subclasses via a special program that allows the user to select whole maps or slices of maps from specific regions of interest and to query by map type. The ability to browse maps stored in GDB or download them in the background was also incorporated into GDB 6.0.

GDB stores many maps of each chromosome, generated by a variety of mapping methods. Users who are interested in a region, such as the neighborhood of a gene or marker, will be able to see all maps that have data in that region, whether or not they contain the desired marker. To support database querying by region of interest, integrated maps have been developed that combine data from all the maps for each chromosome. These are called Comprehensive Maps.

Queries for all loci in a region of interest are processed against the comprehensive maps, thereby searching all relevant maps. Comprehensive maps are also useful for display purposes because they organize the content of a region by class of locus (e.g., gene, marker, clone) rather than by data source. This approach yields a much less complex presentation than an alignment of numerous primary maps. Because such information as detailed orders, order discrepancies between maps, and nonlinear metric relations between maps is not always captured in the comprehensive maps, GDB continues to provide access to aligned displays of primary maps.

A Variety of Searching Strategies

Recognizing the eclectic user commu-nity's need to search data and formulate queries, GDB offers a spectrum of simple to complex search strategies. In addition, direct programming access is available using either GDB's object query language to the Object Broker software layer or standard query language to the underlying Sybase relational database.

Querying by Object Directly from GDB's Home Page

The simplest methods search for objects according to known GDB accession numbers; sequence database–accession numbers; specified names, including wildcard symbols that will automatically match synonyms and primary names; and keywords contained anywhere in the text.

Querying by Region of Interest

A region of interest can be specified using a pair of flanking markers, which can be cytogenetic bands, genes, amplimers (sequence tagged sites), or any other mapped objects. Given a region of interest, the comprehensive maps are searched to find all loci that fall within them. These loci can be displayed in a table, graphically as a slice through a comprehensive map, or as slices through a chosen set of primary maps. A comprehensive map slice shows all loci in the region, including genes, expressed sequence tags (ESTs), amplimers, and clones. A region also can be specified as a neighborhood around a single marker of interest.

Results of queries for genes, amplimers, ESTs, or clones can be displayed on a GDB comprehensive map. Results are spread across several chromosomes displayed in Mapview. A query for all the PAX genes (specified as symbol = PAX* on the gene query form) retrieves genes on multiple chromosomes. Double-clicking on one of these genes brings up detailed gene information via the Web browser.

Querying by Polymorphism

GDB contains a large number of polymorphisms associated with genes and other markers. Queries can be constructed for a particular type of marker (e.g., gene, amplimer, clone), polymorphism (i.e., dinucleotide repeat), or level of heterozygosity. These queries can be combined with positional queries to find, for example, polymorphic amplimers in a region bounded by flanking markers or in a particular chromosomal band. If desired, the retrieved markers can be viewed on a comprehensive map.

Work in Progress

Mapview 2.3

Mapview 2.1, the next generation of the GDB map viewer, was released in March 1997. The latest version, Mapview 2.3, is available in all common computing environments because it is written in the Java programming language. Most important, the new viewer can display multiple aligned maps side by side in the window, with alignment lines indicating common markers in neighboring maps. As before, users can select individual markers to retrieve more information about them from the database.

GDB developers have entered into a collaborative relationship with other members of the bioWidget Consortium so the Java-based alignment viewer will become part of a collection of freely available software tools for displaying biological data (*http://goodman.jax.org/projects/biowidgets/consortium*).

Future plans for Mapview include providing or enhancing the ability to generate manuscript-ready Postscript map images, highlight or modify the display of particular classes of map objects based on attribute values, and requery for additional information.

Variation

Since its inception, GDB has been a repository for polymorphism data, with more than 18,000 polymorphisms now in GDB. A collaboration has been initiated with the Human Gene Mutation Database (HGMD) based in Cardiff, Wales, and headed by David Cooper and Michael Krawczak. HGMD's extensive collection of human mutation data, covering many disease-causing loci, includes sequence-level mutation characterizations. This data set will be included in GDB and updated from HGMD on an ongoing basis. The HGMD team also will provide advice on GDB's representation of genetic variation, which is being enhanced to model mutations and polymorphisms at the sequence level. These modifications will allow GDB to act as a repository for single-nucleotide polymorphisms, which are expected to be a major source of information on human genetic variation in the near future.

Mouse Synteny

Genomic relationships between mouse and man provide important clues regarding gene location, phenotype, and function. One of GDB's goals is to enable direct comparisons between these two organisms, in collaboration with the Mouse Genome Database at Jackson Laboratory. GDB is making additions to its schema to represent this information so that it can be displayed graphically with Mapview. In addition, algorithmic work is under way to use mapping data to automatically identify regions of conserved synteny between mouse and man. These algorithms will allow the synteny maps to be updated regularly. An important application of comparative mapping is the ability to predict the existence and location of unknown human homologs of known, mapped mouse genes. A set of such predictions is available in a report at the GDB Web site, and similar data will be available in the database itself in the spring of 1998.

Collaborations

GDB is a participant in the Genome Annotation Consortium (GAC) project, whose goal is to produce high-quality, automatic annotation of genomic sequences (*http://compbio.ornl.gov/CoLab*). Currently, GDB is developing a prototype mechanism to transition from GDB's Mapview display to the GAC sequence-level browser over common genome regions. GAC also will establish a human genome reference sequence that will be the base against which GDB will refer all polymorphisms and mutations. Ultimately, every genomic object in GDB should be related to an appropriate region of the reference sequence.

Sequencing Progress

The sequencing status of genomic regions now can be recorded in GDB. Based on submissions to sequence databases, GAC will determine genomic regions that have been completed. GDB also will be collaborating with the European Bioinformatics Institute, in conjunction with the international Human Genome Organisation (HUGO), to maintain a single shared Human Sequence Index that will record commitments and status for sequencing clones or regions. As a result, the sequencing status of any region can be displayed alongside other GDB mapping data.

Outreach

The Genome Database continues to seek direct community feedback and interact with the broader science community via various sources:

- International Scientific Advisory Committee meets annually to offer input and advice.
- Quarterly Review Committee confers frequently with the staff to track GDB progress and suggest change.
- HUGO nomenclature, chromosome, and other editorial committees have specialized functions within GDB, providing official names and consensus maps and ensuring the high quality of GDB's content.

Copies of GDB are available worldwide from ten mirror sites (nodes), and GDB staff members meet annually with node managers.

National Center for Genome Resources

Genome Sequence DataBase
1800 Old Pecos Trail, Suite A
Santa Fe, NM 87505

Peter Schad, Vice-President
Bioinformatics and Biotechnology
505/995-4447, Fax: -4432
cnc@ncgr.org

Carol Harger
GSDB Manager
505/982-7840, Fax: -7690
cah@ncgr.org

http://www.ncgr.org

The National Center for Genome Resources (NCGR) is a not-for-profit organization created to design, develop, support, and deliver resources in support of public and private genome and genetic research. To accomplish these goals, NCGR is developing and publishing the Genome Sequence DataBase (GSDB) and the Genetics and Public Issues (GPI) program.

NCGR is a center to facilitate the flow of information and resources from genome projects into both public and private sectors. A broadly based board of governors provides direction and strategy for the center's development.

NCGR opened in Santa Fe in July 1994, with its initial bioinformatics work being developed through a cooperative 5-year agreement with the Department of Energy funded in July 1995. Committed to serving as a resource for all genomic research, the center works collaboratively with researchers and seeks input from users to ensure that tools and projects under development meet their needs.

Genome Sequence DataBase

GSDB is a relational database that contains nucleotide sequence data and its associated annotation from all known organisms (*http://www.ncgr.org/gsdb*). All data are freely available to the public. The major goals of GSDB are to provide the support structure for storing sequence data and to furnish useful data-retrieval services.

GSDB adheres to the philosophy that the database is a "community-owned" resource that should be simple to update to reflect new discoveries about sequences. A corollary to this is GSDB's conviction that researchers know their areas of expertise much better than a database curator and, therefore, they should be given ownership and control over the data they submit to the database. The true role of the GSDB staff is to help researchers submit data to and retrieve data from the database.

GSDB Enhancements

During 1996, GSDB underwent a major renovation to support new data types and concepts that are important to genomic research. Tables within the database were restructured, and new tables and data fields were added. Some key additions to GSDB include the support of data ownership, sequence alignments, and discontiguous sequences.

The concept of data ownership is a cornerstone to the functioning of the new GSDB. Every piece of data (e.g., sequence or feature) within the database is owned by the submitting researcher, and changes can be made only by the data owner or GSDB staff. This implementation of data ownership provides GSDB with the ability to support community (third-party) annotation—the addition of annotation to a sequence by other community researchers.

A second enhancement of GSDB is the ability to store and represent sequence alignments. GSDB staff has been constructing alignments to several key sequences including the env and pol (reverse transcriptase) genes of the HIV genome, the complete chromosome VIII of Saccharomyces cerevisiae, and the complete genome of Haemophilus influenzae. These alignments are useful as possible sites of biological interest and for rapidly identifying differences between sequences.

A third key GSDB enhancement is the ability to represent known relationships of order and distance between separate individual pieces of sequence. These sets of sequences and their relative positions are grouped together as a single discontiguous sequence. Such a sequence may be as simple as two primers that define the ends of a sequence tagged site (STS), it may comprise all exons that are part of a single gene, or it may be as complex as the STS map for an entire chromosome.

GSDB staff has constructed discontigu-ous sequences for human chromosomes 1 through 22 and X that include markers from Massachusetts Institute of Technology–Whitehead Institute STS maps and from the Stanford Human Genome Center. The set of 2000 STS markers for chromosome X, which were mapped recently by Washington University at St. Louis, also have been added to chromosome X. About 50 genomic sequences have been added to the chromosome 22 map by determining their overlap with STS markers. Genomic sequences are being added to all the chromosomes as their overlap with the STS markers is determined. These discontigu-ous sequences can be retrieved easily and viewed via their sequence names using the GSDB Annotator. Sequence names follow the format of HUMCHR#MP, where # equals 1 through 22 or X.

GSDB staff also has utilized discontigu-ous sequences to construct maps for maize and rice. The maize discontiguous sequences were constructed using markers from the University of Missouri, Columbia. Markers for the rice discontiguous sequence were obtained from the Rice Genome Database at Cornell University and the Rice Genome Research Project in Japan.

New Tools

As a result of the major GSDB renovation, new tools were needed for submitting and accessing database data. Annotator was developed as a graphical interface that can be used to view, update, and submit sequence data (*http://www.ncgr.org/gsdb/beta.html*). Maestro, a Web-based interface, was developed to assist researchers in data retrieval (*http://www.ncgr.org/gsdb/maestrobeta.html*). Although both these tools currently are available to researchers, GSDB is continuing development to add increased capabilities.

Annotator displays a sequence and its associated biological information as an image, with the scale of the image adjustable by the user. Additional information about the sequence or an associate biological feature can be obtained in a pop-up window. Annotator also allows a user to retrieve a sequence for review, edit existing data, or add annotation to the record. Sequences can be created using Annotator, and any sequences created or edited can be saved either to a local file for later review and further editing or saved directly to the database.

Correct database structures are important for storing data and providing the research community with tools for searching and retrieving data. GSDB is making a concerted effort to expand and improve these services. The first generation of the Maestro query tool is available from the GSDB Web pages. Maestro allows researchers to perform queries on 18 different fields, some of which are queryable only through GSDB, for example, D segment numbers from the Genome Database at Johns Hopkins University in Baltimore.

Additionally, Maestro allows queries with mixed Boolean operators for a more refined search. For example, a user may wish to compare relatively long mouse and human sequences that do not contain identified coding regions. To obtain all sequences meeting these criteria, the scientific name field would be searched first for "Mus musculus" and then for "Homo sapiens" using the Boolean term "OR." Then the sequence-length filter could be used to refine the search to sequences longer than 10,000 base pairs. To exclude sequences containing identified coding-region features, the "BUT NOT" term can be used with the Feature query field set equal to "coding region."

With Maestro, users can view the list of search matches a few at a time and retrieve more of the list as needed. From the list, users can select one or several sequences according to their short descriptions and review or download the sequence information in GIO, FASTA, or GSDB flatfile format.

Future Plans

Although most pieces necessary for operation are now in place, GSDB is still improving functionality and adding enhancements. During the next year GSDB, in collaboration with other researchers, anticipates creating more discontiguous sequence maps for several model organisms, adding more functionality to and providing a Web-based submission tool and tool kit for creating GIO files.

Microbial Genome Web Page

NCGR also maintains informational Web pages on microbial genomes. These pages, created as a community reference, contain a list of current or completed eubacterial, Archaeal, and eukaryotic genome sequencing projects. Each main page includes the name of the organism being sequenced, sequencing groups involved, background information on the organism, and its current location on the Carl Woese Tree of Life. As the Microbial Genome Project progresses, the pages will be updated as appropriate.

Genetics and Public Issues Program

GPI serves as a crucial resource for people seeking information and making decisions about genetics or genomics (*http://www.ncgr.org/gpi*). GPI develops and provides information that explains the ethical, legal, policy, and social relevance of genetic discoveries and applications.

To achieve its mission, GPI has set forth three goals: (1) preparation and development of resources, including careful delineation of ethical, legal, policy, and social issues in genetics and genomics; (2) dissemination of genetic information targeted to the public, legal and health professionals, policymakers, and decision makers; and (3) creation of an information network to facilitate interaction among groups.

GPI delivers information through four primary vehicles: online resources, conferences, publications, and educational programs. The GPI program maintains a continually evolving World Wide Web site containing a range of material freely accessible over the Internet.

Index to Principal and Coinvestigators Listed in Abstracts

A

Adams, Mark D. 8
Adamson, Doug 6
Adamson, Anne E. 59
Agarwal, Pankaj 41
Aksenov, N.D. 26
Albertson, Donna 7
Allison, David 19
Allman, Steve L. 1
Anderson, Holt 70
Anderson, J. Clarke 70
Annas, George J. 69
Apostolou, Sinoula 68
Apsell, Paula 69
Arenson, A. 23
Arlinghaus, Heinrich F. 67, 70
Arman, Inga P. 67
Ashworth, Linda 28
Athwal, Raghbir S. 67
Aytay, Saika 70

B

Baker, Diane 69
Baker, Elizabeth 68
Baker, Mark E. 67
Banerjee, Subrata 30
Baranova, A.V. 30
Barber, William M. 68
Barker, David L. 70
Barsky, V. 10
Bashiardes, Evy 30
Baumes, Susan 27
Bavikin, S. 10
Bayne, Peter 70
Beeson, Diane 48
Belikov, S.V. 22
Benner, W.H. 1
Binder, Matt 53
Birren, B. 68
Blatt, Robin J.R. 53
Blinov, Vladimir M. 67
Boitsov, Alexandre S. 19
Boitsov, Stepan A. 19
Bonaldo, Maria de Fatima 27
Boughton, Ann 55
Bradley, J.-C. 67
Branscomb, Elbert 28
Bremer, Meire 68
Brennan, Thomas M. 67
Bridgers, Michael A. 68
Briley, J. David 13
Brody, Linnea 68
Bronstein, Irena 70
Brown, Gilbert M. 67
Brown, Henry T. 68
Browne, Murray 59
Bruce, J. E. 15
Bruce, James E. 14
Bugaeva, Elena 24
Bulger, Ruth E. 69
Bumgarner, Roger 68
Buneman, Peter 39
Burbee, Dave 4, 5
Burks, Christian 68
Butler-Loffredo, Laura-Li 3

C

Cacheiro, Nestor 29
Callen, David F. 68
Cantor, Charles R. 19
Capron, Alex 69
Carlson, Charles C. 45, 69
Carrano, Anthony V. 68
Cartwright, Peter 6
Carver, Ethan 28, 29
Casey, Denise K. 59
Catanese, Joe 20
Chait, Brian 14
Chang, Huan-Tsung 17
Chedd, Graham 45, 69
Chen, Chira 20
Chen, Chung-Hsuan 1
Chen, Ed 69
Chen, I-Min A. 36
Chen, X-N. 68
Cherkauer, Kevin 69
Chetverin, Alexander B. 68
Chikaev, N.A. 67
Chinault, A.C. 23
Chittenden, Laura 29
Chou, Chau-Wen 17
Chou, Hugh 41
Church, George 2
Churchill, Gary 68
Cinkosky, Michael J. 68
Cobbs, Archie 69
Collins, Colin 7
Collins, Debra L. 45
Conn, Lane 46
Cozza, S. 37
Cram, L.S. 26
Crandall, Lee A. 69
Craven, Mark 69
Crkvenjakov, Radomir 67, 68
Cuddihy, D. 37
Culiat, Cymbeline 29
Cytron, Ron 41

DOE Human Genome Progress Report, Part 2: Index

D

Davidson, Jack B. 67
Davidson, Jeff 47
Davidson, Susan B., 39
Davies, Chris 4, 5
Davis, Sharon 47
Davison, Daniel 33
de Jong, P. 68
de Jong, Pieter 2
de Jong, Pieter J. 20
Denton, M. Bonner 67
Dettloff, Wayne 70
Devin, Alexander B. 67
Di Sera, Leonard 6
Doggett, Norman A. 68
Dogruel, David 17
Doktycz, Mitch 19
Dovichi, Norman 3
Doyle, Johannah 28, 29
Drmanac, Radoje 67, 68
Drmanac, Snezana 67
Dunn, Diane 6
Dunn, John J. 3, 4
Durkin, Scott 68
Duster, Troy 48
Dyer, Joshua P. 64

E

Eadline, Douglas J. 70
Earle, Colin W. 67
Efimenko, Irina G. 67
Egenberger, Laurel 54
Eichler, E.E. 23
Einstein, J. Ralph 42
Eisenberg, Rebecca S. 48
Enukashvily, Natella 24
Evans, Glen A. 4, 5, 67

F

Fader, Betsy 69
Fallon, Lara 67
Ferguson, F. Mark 6
Ferrell, Thomas L. 67
Fickett, James W. 68
Fields, Christopher A. 69
Filipenko, M.L. 67
Firulli, B.A. 23
Flatley, Jay 70
Florentiev, V.L. 5
Fockler, Carita 68
Fodor, Stephen P. A. 70
Fondon, Trey 4
Foote, Robert S. 67
Franklin, Terry 4, 5
Frengen, Eirik 20
Fresco, Jacques R. 21
Friedman, B. Ellen 51, 52
Friedman, Claudette Cyr 69
Fullarton, Jane E. 69
Fung, Eliza 17

G

Gaasterland, Terry 38
Garner, Harold R. (Skip) 4, 5
Gath, Tracy 54
Generoso, Walderico 29
Gerwehr, S. 68
Gesteland, Raymond F. 6
Gibbs, R.A. 23
Glantz, Leonard H. 69
Glazer, Alexander N. 9
Glazkova, Dina V. 67
Gnirke, Andreas 68
Golumbeski, George 70
Goodman, Nathan 33
Goodman, Stephen 49
Graves, M. 23
Graves, Mark 34
Gray, Joe 7
Gregory, Paula 69
Griffith, Jeffrey K. 12
Grosz, Michael 30
Gu, Y. 23
Guan, Xiaojun 42
Guan, Xiaoping 20
Guilfoyle, Richard A. 13
Gusfield, Dan 69

H

Hahn, Peter 68
Hahner, Lisa 4
Hartman, John R. 70
Hartnett, Jim 70
Hauser, Loren 42, 44
Haussler, David 34
Hawe, William P. 67
Hawkins, Trevor 8
Hempfner, Philip E. 68
Henderson, Margaret 70
Hofstadler, S. A. 15
Holmes, Linda 59
Hood, Leroy 8, 52, 69
Hooper, Herbert H. 70
Hopkins, Janet A. 68
Horton, Paul 69
Hoyt, Peter 19
Hozier, John 68
Hubert, R. 68
Hughey, Richard 34
Hung, Lydia 70
Hunkapiller, Tim 69

I

Ijadi, Mohamad 68
Il'icheva, I.A. 5
Imara, Mwalimu 69
Ioannou, Panayotis A. 20, 30
Ivanovich, M.A. 67
Iwasaki, R. 37

J

Jackson, Cynthia L. 67
Jacobson, K. Bruce 1, 67
Jaklevic, J.M. 1
Jantsen, E.I. 67
Jefferson, Margaret C. 50
Jelenc, Pierre 27
Jessee, Joel 20
Johnson, Marion D., III 21
Jurka, Jerzy 34

K

Kamashev, D.E. 22
Kao, Fa-Ten 21
Kapanadze, B.I. 30
Karger, Barry L. 9
Karp, Richard 69
Karp, Richard M. 69
Karplus, Kevin 34
Karpov, V.L. 22
Kass, Judy 54
Kaur, G. Pal 67
Kel, A.E. 35
Kel, O.V. 35
Keller, Richard 67
Khan, Akbar S. 68
Kim, Joomyeong 28
Kim, U-J. 68
Kim, Ung-Jin 26, 27
Kimball, Alvin 6
Klopov, N.V. 26
Knight, Jim 69
Knoche, Kimberly 70
Knoppers, Bartha 69
Knuth, Mark W. 63
Kolchanov, N.A. 35
Korenberg, J.R. 68
Korenberg, Julie 20
Korenberg, Julie R. 22
Kozman, Helen 68
Krasnykh, Viktor N. 67
Krone, Jennifer 17
Kupfer, Ken 5
Kwok, Pui-Yan 68

L

Labat, Ivan 67, 68
Lai, Tran N. 68
Lander, E. 68
Lane, Michael J. 68
Lane, Sharon A. 68
Lantos, John 50
Larimer, Frank W. 67
Larson, Susan 38
Lawler, Gene 69
Lazareva, Betty 68
Legchilina, Svetlana P. 67
Lennon, Greg 29
Leone, Joseph 64
Lessick, Mira 50
Lever, David C. 67
Lewis, Kathy 17
Li, Qingbo 17
Lim, Hwa A. 69
Lim, Regina 68
Lobov, Ivan 24
Lockett, Steven 7
Lu, J. 23
Lu, Xiandan 17
Luchina, N.N. 25
Lukjanov, Dmitry 24
Lvovsky, Lev 16
Lysov, Y. 10

M

MacConnell, William P. 64
MacDonell, Michael T. 70
Maglott, Donna R. 68
Mahowald, Mary B. 50
Mallison, M. 37
Maltsev, Natalia 38
Mann, Janice 55
Manning, Ruth Ann 64
Mansfield, Betty K. 59
Manske, Charles L. 70
Mark, Hon Fong L. 67
Markowitz, Victor M. 36
Marks, Andy 6
Marr, T. 37
Martin, Sheryl A. 59
Martin, Chris S. 70
Mathies, Richard A. 9
Matis, Sherri 42
Matveev, Ivan 24
McAllister, Douglas 70
McAllister, Douglas J. 70
McInerney, Joseph D. 51, 52
Metzger, M. 23
Micikas, Lynda B. 52
Micklos, David A. 70

Mills, Marissa D. 59
Milosavljevic, Aleksandar 68
Mirzabekov, Andrei 10
Mishin, V.P. 67
Mitchell, S. 68
Moore, Stefan 69
Mosley, Ray E. 69
Moss, Robert 50
Moyzis, Robert K. 12
Muddiman, David C. 14, 15
Mulley, John C. 68
Munn, Maureen M. 52
Mural, Richard 44
Mural, Richard J. 42
Muravlev, A.I. 67
Murphy, Declan 69
Murphy, Kevin 69
Muzny, D.M. 23
Myers, Gene 38

N

Nancarrow, Julie 68
Natowicz, Marvin 69
Nelson, D. L. 23
Nelson, Debra 68
Nelson, Randall 17
Newman, Cathy D. 70
Nguyen, Tuyen 64
Nicholls, Robert 29
Nickerson, Deborah A. 68
Nierman, William C. 68
Noordewier, Michiel O. 69
Noya, D. 68

O

Olenina, Ludmilla V. 67
Olesen, Corinne E. M., 70
Oliver, Tammy 4
Olson, Maynard 68
Olson, Maynard V. 52
Orpana, Arto K. 68
Oskin, Boris V. 19
Ostrander, Elaine A. 68
Overbeek, Ross 38
Overton, G. Christian 39, 41
Overton, G.C. 35

P

Page, George 69
Pecherer, Robert M. 68
Petrov, Sergey 42, 44
Pevzner, Pavel A. 40
Pfeifer, Gerd P. 67
Phillips, Hilary A. 68
Phoenix, David 69
Pietrzak, Eugenia 20
Pinkel, Daniel 7
Pirrung, Michael C. 67
Podgornaya, Olga 24
Podkolodnaya, O.A. 35
Polanovsky, O.L. 25
Poletaev, A.I. 26
Polymeropoulos, Mihael H. 68
Porter, Kenneth W. 13
Pratt, Lorien 69
Preobrazhenskaya, O.V. 22
Probst, Shane 4, 5

R

Radspinner, David A. 67
Raja, Mugasimangalam 16
Randesi, Matthew 4
Reed, C. 37
Reilly, Philip 69
Reilly, Philip J. 53
Resenchuk, Sergei M. 67
Reshetin, Anton O. 19
Richards, Robert I. 68
Richterich, Peter 70
Rider, Michelle 30
Riggs, Arthur D. 67
Roche, Patricia A. 69
Romaschenko, A.G. 35
Ross, Lainie Friedman 50
Roszak, Darlene B. 70
Roth, E.J. 23
Rozen, Steve 33
Ruano, Gualberto 63
Rutledge, Joe 29

S

Sachleben, Richard A. 67
Sachs, Greg 50
Sainz, Jesus 68
Salit, J. 37
Sandakhchiev, Lev S. 67
Sandhu, Arbansjit K. 67
Schageman, Jeff 5
Schimke, R. Neil 45
Schurtz, Tony 6
Schwerin, Noel 45
Scott, Bari 53
Searls, David B. 41
Selkov, Evgeni 38
Selman, Susanne 70
Semov, A.B. 30
Serpinsky, Oleg I. 67
Sesma, Mary Ann 50
Sgro, Peichen H. 68
Shah, Manesh 42, 44
Shannon, Mark 28

Sharpe, Elizabeth 69
Shatrova, A.N. 26
Shavlik, Jude W. 69
Shaw, Barbara Ramsay 13
Shchelkunov, Sergei N. 67
Shchyolkina, A.K. 5
Shen, Y. 23
Shick, V. 10
Shizuya, H. 68
Shizuya, Hiroaki 26, 27
Shuey, Steven W. 67
Siciliano, Michael J. 68
Sikela, James M. 68
Silva, J. 68
Simon, M. 68
Simon, Melvin 8
Simon, Melvin I. 26, 27
Sivila, Randy F. 64
Smirnova, Marina E. 67
Smirnova, V.V. 67
Smith, Cassandra L. 68
Smith, Lloyd M. 13, 14, 67
Smith, Randall 33
Smith, Richard D. 14, 15
Soane, David S. 70
Soares, Marcelo Bento 27
Soderlund, Carol A. 69
Solomon, David L. 70
Sonkin, Dina 16
Sorenson, Doug 68
Sosa, Maria 54
Spejewski, Eugene 59
Spengler, Sylvia 55
Spengler, Sylvia J. 54, 60
States, David J. 41
Stavropoulos, Nick 68
Stein, Lincoln 33
Stelling, Paul 69
Stepchenko, A.G. 25
Stevens, Tamara J. 68
Stormo, Gary D. 69
Stubbs, Lisa 28, 29
Studier, F. William 3, 4
Sudar, Damir 7
Sulimova, G.E. 30
Sun, Tian-Qiang 30
Sun, Z. 68
Sutherland, Grant R. 68
Sutherland, Robert D. 68
Sze, Sing Hoi 40

T

Tabor, Stanley 16, 67
Thilman, Jude 53
Thonnard, Norbert 67
Thundat, Thomas G. 67
Thundat, Tom 19
Timms, K 23
Timofeev, E.N. 5
Tobin, Sara L. 55
Totmenin, Alexei V. 67
Towell, Geoffrey 69
Tracy, A. 37
Trask, Barbara 68
Trask, Barbara J. 68
Trottier, Ralph W. 69
Troup, Charles D. 68
Tsybenko, S. Yu 5

U

Uberbacher, Edward 44
Uberbacher, Edward C. 42
Udseth, Harold R. 14
Ulanovsky, Levy 16

V

van den Engh, Ger 68
Verp, Marion 50
Vos, Jean-Michel H. 30

W

Wahl, Geoffrey 68
Walkowicz, Mitchell 29
Wang, Denan 68
Wang, Lushen 69
Wang, Min 30
Warmack, Bruce 19
Warmack, Robert J. 67
Wassom, John S. 59
Waterman, Michael 69
Weier, Heinz-Ulrich 7
Weinberger, Laurence 47
Weiss, Robert B. 6
Wentland, M.A. 23
Wertz, Dorothy C. 53
Westin, Alan F. 69
Whitmore, Scott A. 68
Whitsitt, Andrew 69
Wilcox, Andrea S. 68
Williams, Peter 17
Williams, Walter 61
Wingender, E. 35
Witkowski, Jan 70
Wong, Gane 68
Woychik, Richard P. 67
Wright, Gary 29
Wright, James 61
Wu, Chenyan 20
Wu, J. 23
Wu, X. 68
Wyrick, Judy M. 59

X

Xu, Ying 42

Y

Yankovsky, N.K. 30
Yantis, Bonnie C. 68
Yershov, G. 10
Yeung, Edward S. 17
Yoshida, Kaoru 68
Yu, Jun 68
Yust, Laura N. 59

Z

Zenin, V.V. 26
Zhao, Baohui 20
Zoghbi, H.Y. 23
Zorn, Manfred 7
Zorn, Manfred D. 44
Zweig, Franklin M. 56

SCIENTIFIC
AMERICAN

Discovering Genes for New Medicines

by William A. Haseltine

SCIENTIFIC
AMERICAN
MARCH 1997
VOL. 276, NO. 3 PP. 92–97

Copyright © 1997 by Scientific American, Inc. All rights reserved. Printed in the U.S.A. No part of this reprint may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. The trademark and tradename "SCIENTIFIC AMERICAN" and the distinctive logotype pertaining thereto are the sole property of and are registered under the name of Scientific American, Inc. Page numbers and internal references may vary from those in original issue.

The SciDex® Electronic Index, of Scientific American feature articles since 1948, is available at low cost. To order, call 1-800-777-0444. Outside of the U.S. and Canada, write to: Scientific American, Attn: SciDex, 415 Madison Avenue, New York, NY 10017-1111.

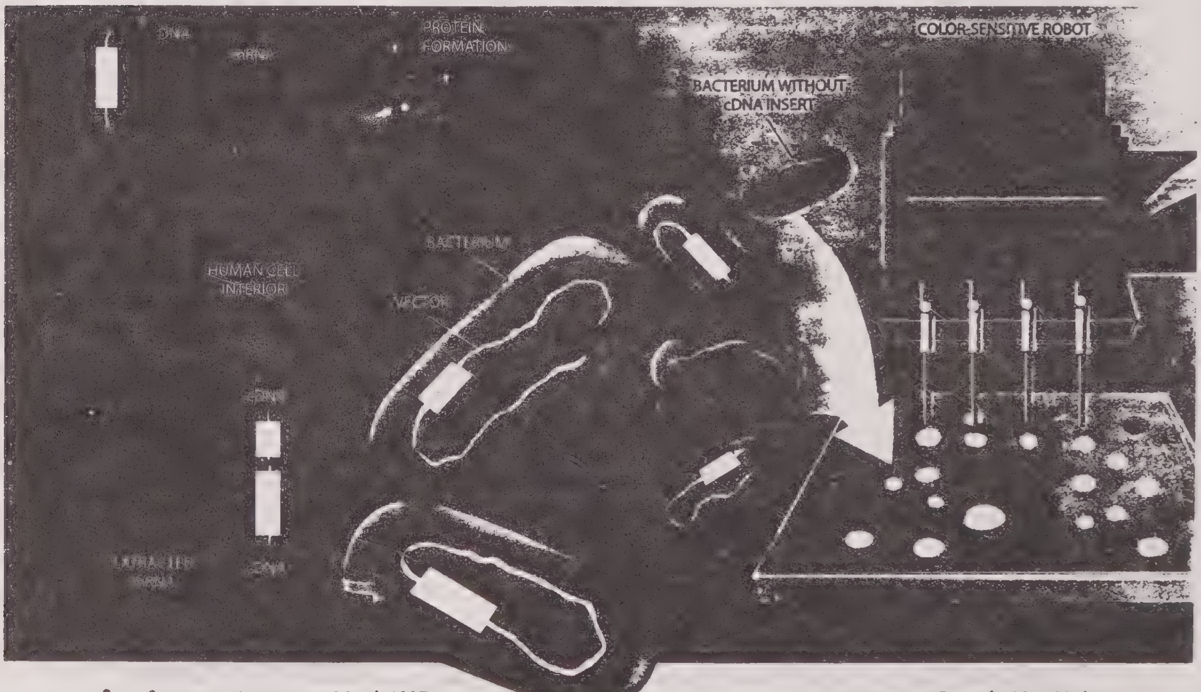
Discovering Genes for New Medicines

By identifying human genes involved in disease, researchers can create potentially therapeutic proteins and speed the development of powerful drugs

by William A. Haseltine

Most readers of this magazine are probably familiar with the idea of a gene as something that transmits inherited traits from one generation to the next. Less well appreciated is that malfunctioning genes are deeply involved in most diseases, not only inherited ones. Cancer, atherosclerosis, osteoporosis, arthritis and Alzheimer's disease, for example, are all characterized by specific changes in the activities of genes. Even infectious disease usually provokes the activation of identifiable genes in a patient's immune system. Moreover, accumulated damage to genes from a lifetime of exposure to ionizing radiation and injurious chemicals probably underlies some of the changes associated with aging.

A few years ago I and some like-minded colleagues decided that knowing where and when different genes are switched on in the human body would lead to far-reaching advances in our ability to predict, prevent, treat and cure disease. When a gene is active, or as a geneticist would say, "expressed," the sequence of the chemical units, or bases, in its DNA is used as a blueprint to produce a specific protein. Proteins direct, in various ways, all of a cell's functions. They serve as structural components, as catalysts that carry out the multiple chemical processes of life and as control elements that regulate cell reproduction, cell specialization and physiological activity at all levels. The development of a human from fertilized egg to mature adult is, in fact, the consequence of an



orderly change in the pattern of gene expression in different tissues.

Knowing which genes are expressed in healthy and diseased tissues, we realized, would allow us to identify both the proteins required for normal functioning of tissues and the aberrations involved in disease. With that information in hand, it would be possible to develop new diagnostic tests for various illnesses and new drugs to alter the activity of affected proteins or genes. Investigators might also be able to use some of the proteins and genes we identified as therapeutic agents in their own right. We envisaged, in a sense, a high-resolution description of human anatomy descending to the molecular level of detail.

It was clear that identifying all the expressed genes in each of the dozens of tissues in the body would be a huge task. There are some 100,000 genes in a typical human cell. Only a small proportion of those genes (typically about 15,000) is expressed in any one type of cell, but the expressed genes vary from one cell type to another. So looking at just one or two cell types would not reveal the genes expressed in the rest of the body. We would also have to study tissues from all the stages of human development. Moreover, to identify the changes in gene expression that contribute to sickness, we would have to analyze diseased as well as healthy tissues.

Technological advances have provided a way to get the job done. Scientists can now rapidly discover which genes are expressed in any given tissue. Our strategy has proved the quickest way to identify genes of medical importance.

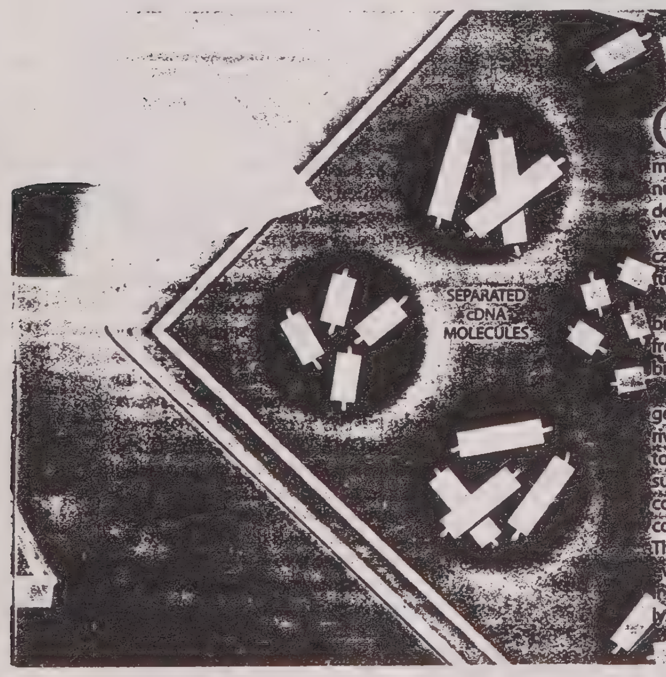
Take the example of atherosclerosis. In this common condition, a fatty substance called plaque accumulates inside arteries, notably those supplying the heart. Our strategy enables us to generate a list of genes expressed in normal arteries, along with a measure of the level of expression of each one. We can then compare the list with one derived from patients with atherosclerosis. The difference between the lists corresponds to the genes (and thus the proteins) involved in the disease. It also indicates how much the genes' expression has been increased or decreased by the illness. Researchers can then make the human proteins specified by those genes.

Once a protein can be manufactured in a pure form, scientists can fairly easily fashion a test to detect it in a patient. A test to reveal overproduction of a protein found in plaque might expose early signs of atherosclerosis, when better options exist for treating it. In addition, pharmacologists can use pure proteins to help them find new drugs. A chemical that inhibited production of a protein found in plaque might be considered as a drug to treat atherosclerosis.

Our approach, which I call medical genomics, is somewhat outside the mainstream of research in human genetics. A great many scientists are involved in the Human Genome Project, an international collaboration devoted to the discovery of the complete sequence of the chemical bases in human DNA. (All the codes in DNA are constructed from an alphabet consisting of just four bases.) That information will be important for studies of gene action and evolution and will particularly benefit research on inherited diseases. Yet the genome project is not the fastest way to discover genes, because most of the bases that make up DNA actually lie outside genes. Nor will the project pinpoint which genes are involved in illness.

In 1992 we created a company, Human Genome Sciences (HGS), to pursue our vision. Initially we conducted the work as a collaboration between HGS and the Institute for Genomic Research, a not-for-profit organization that HGS supports; the institute's director, J. Craig Venter, pioneered some of the key ideas in genomic research. Six months into the collaboration, SmithKline Beecham,

JENNIFER C. CHRISTIANSEN



How to Make and Separate cDNA Molecules

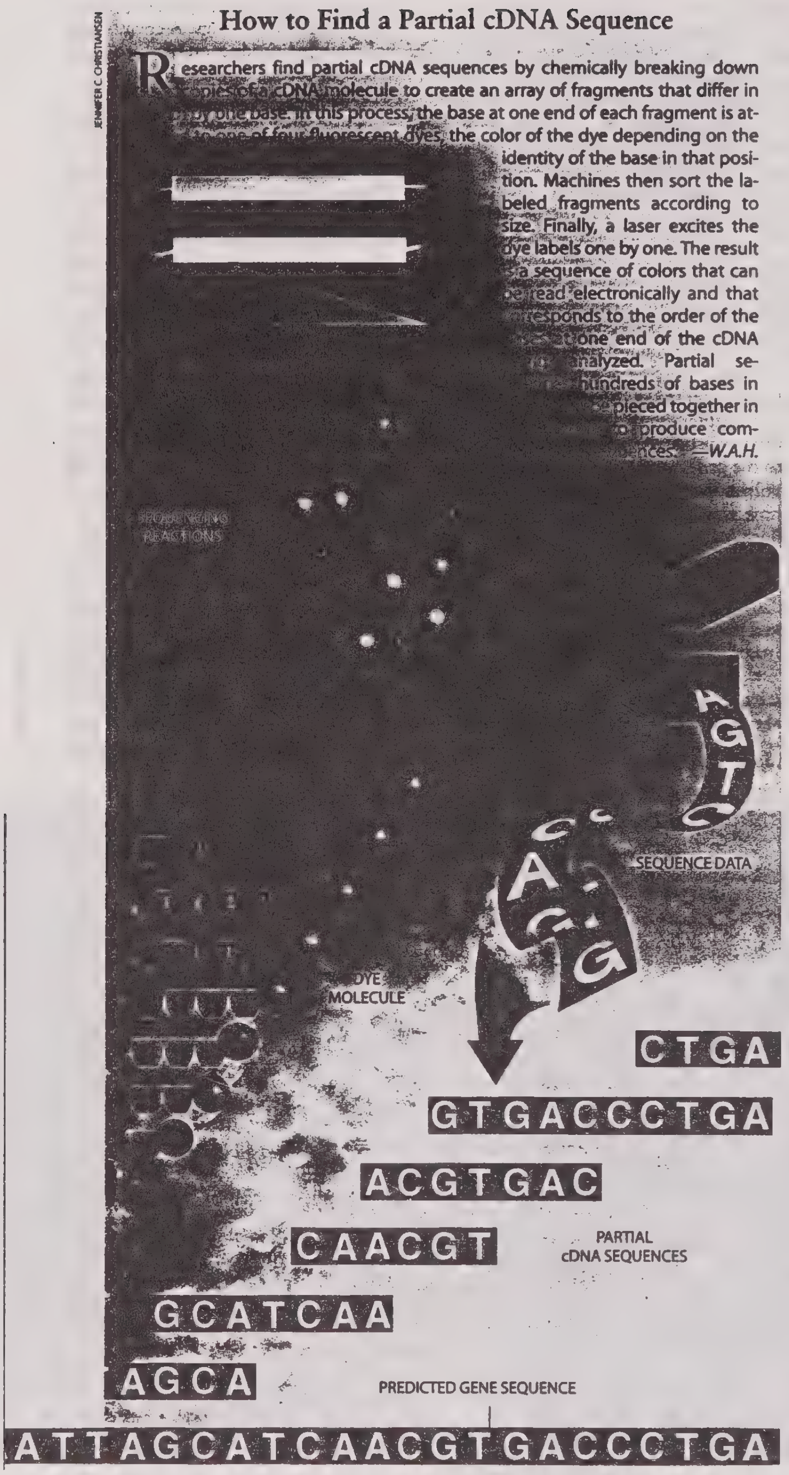
Cells use messenger RNA to make protein. We discover genes by making complementary DNA (cDNA) copies of messenger RNA. First we have to clone and produce large numbers of copies of each cDNA, so there will be enough to determine its constituent bases. Molecular biologists have developed ways to insert cDNA into specialized DNA loops, called vectors, that can reproduce inside bacterial cells. A mixture of cDNAs from a given tissue is called a library.

Researchers at HGS have now prepared human cDNA libraries from almost all normal organs and tissues, as well as from many that are diseased. To make multiple copies of a library, we add it to bacteria that take up the vectors.

All the bacteria are then spread out on a plate of nutrient gel and allowed to grow into colonies, so that each colony derives from a single bacterium. Next we use a robot that can automatically spot and pick off the gel those colonies that did successfully acquire a cDNA. The robot accomplishes this by color. The vectors we use are designed so that if they fail to combine with a cDNA insert, they produce a blue pigment. The robot, which picks as many as 10,000 colonies of bacteria every day, identifies those containing human cDNA by avoiding blue ones. The cDNA from each picked colony, now in analyzable quantities, is then robotically purified. *—W.A.H.*

How to Find a Partial cDNA Sequence

Researchers find partial cDNA sequences by chemically breaking down copies of a cDNA molecule to create an array of fragments that differ in only by one base. In this process, the base at one end of each fragment is at[illegible] to one of four fluorescent dyes, the color of the dye depending on the identity of the base in that position. Machines then sort the labeled fragments according to size. Finally, a laser excites the dye labels one by one. The result is a sequence of colors that can be read electronically and that corresponds to the order of the bases at one end of the cDNA [illegible] analyzed. Partial se[illegible] hundreds of bases in [illegible] pieced together in [illegible] to produce com[illegible]ences. —*W.A.H.*



one of the world's largest pharmaceutical companies, joined HGS in the effort. After the first year, HGS and SmithKline Beecham continued on their own. We were joined later by Schering-Plough, Takeda Chemical Industries in Japan, Merck KGaA in Germany and Synthelabo in France.

Genes by the Direct Route

Because the key to developing new medicines lies principally in the proteins produced by human genes, rather than the genes themselves, one might wonder why we bother with the genes at all. We could in principle analyze a cell's proteins directly. Knowing a protein's composition does not, however, allow us to make it, and to develop medicines, we must manufacture substantial amounts of proteins that seem important. The only practical way to do so is to isolate the corresponding genes and transplant them into cells that can express those genes in large amounts.

Our method for finding genes focuses on a critical intermediate product created in cells whenever a gene is expressed. This intermediate product is called messenger RNA (mRNA); like DNA, it consists of sequences of four bases. When a cell makes mRNA from a gene, it essentially copies the sequence of DNA bases in the gene. The mRNA then serves as a template for constructing the specific protein encoded by the gene. The value of mRNA for research is that cells make it only when the corresponding gene is active. Yet the mRNA's base sequence, being simply related to the sequence of the gene itself, provides us with enough information to isolate the gene from the total mass of DNA in cells and to make its protein if we want to.

For our purposes, the problem with mRNA was that it can be difficult to handle. So we in fact work with a surrogate: stable DNA copies, called complementary DNAs (cDNAs) of the mRNA molecules. We make the cDNAs by simply reversing the process the cell uses to make mRNA from DNA.

The cDNA copies we produce this way are usually replicas of segments of mRNA rather than of the whole molecule, which can be many thousands of bases long. Indeed, different parts of a gene can give rise to cDNAs whose common origin may not be immediately apparent. Nevertheless, a cDNA containing just a few thousand bases still preserves its parent gene's unique signature.

That is because it is vanishingly unlikely that two different genes would share an identical sequence thousands of bases long. Just as a random chapter taken from a book uniquely identifies the book, so a cDNA molecule uniquely identifies the gene that gave rise to it.

Once we have made a cDNA, we can copy it to produce as much as we want. That means we will have enough material for determining the order of its bases. Because we know the rules that cells use to turn DNA sequences into the sequences of amino acids that constitute proteins, the ordering of bases tells us the amino acid sequence of the corresponding protein fragment. That sequence, in turn, can be compared with the sequences in proteins whose structures are known. This maneuver often tells us something about the function of the complete protein, because proteins containing similar sequences of amino acids often perform similar tasks.

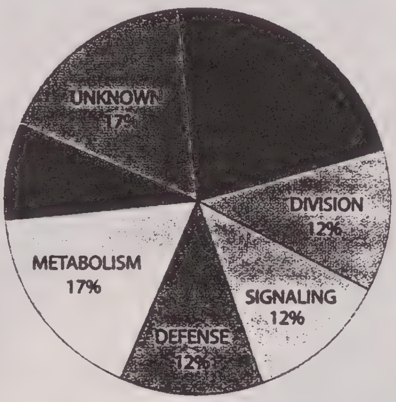
Analyzing cDNA sequences used to be extremely time-consuming, but in recent years biomedical instruments have been developed that can perform the task reliably and automatically. Another development was also necessary to make our strategy feasible. Sequencing equipment, when operated on the scale we were contemplating, produces gargantuan amounts of data. Happily, computer systems capable of handling the resulting megabytes are now available, and we and others have written software that helps us make sense of this wealth of genetic detail.

Assembling the Puzzle

Our technique for identifying the genes used by a cell is to analyze a sequence of 300 to 500 bases at one end of each cDNA molecule. These partial cDNA sequences act as markers for genes and are sometimes referred to as expressed sequence tags. We have chosen this length for our partial cDNA sequences because it is short enough to analyze fairly quickly but still long enough to identify a gene unambiguously. If a cDNA molecule is like a chapter from a book, a partial sequence is like the first page of the chapter—it can identify the book and even give us an idea what the book is about. Partial cDNA sequences, likewise, can tell us something about the gene they derive from. At HGS, we produce about a million bases of raw sequence data every day.

Our method is proving successful: in less than five years we have identified thousands of genes, many of which may play a part in illness. Other companies and academic researchers have also initiated programs to generate partial cDNA sequences.

HGS's computers recognize many of the partial sequences we produce as deriving either from one of the 6,000 genes researchers have already identified by other means or from a gene we have previously found ourselves. When we cannot definitely assign a newly generated partial sequence to a known gene, things get more interesting. Our computers then scan through our databases as well as public databases to see whether the new partial sequence overlaps something someone has logged before. When we find a clear overlap, we piece together the overlapping partial sequences into ever lengthening segments called contigs. Contigs correspond, then, to incomplete sequences we infer to be present somewhere in a parent gene. This process is somewhat analogous to fishing out the phrases "a midnight dreary, while I pondered" and "while I pondered, weak and weary/Over many a...volume" and combining them into a fragment recognizable as part of Edgar Allan Poe's "The Raven."



JENNIFER C. CHRISTIANSEN

PERCENTAGE OF GENES devoted to each of the major activities in the typical human cell has been deduced from a study of 150,000 partial sequences. Similarities with human or other genes of known function were used to assign provisional categories of activity.

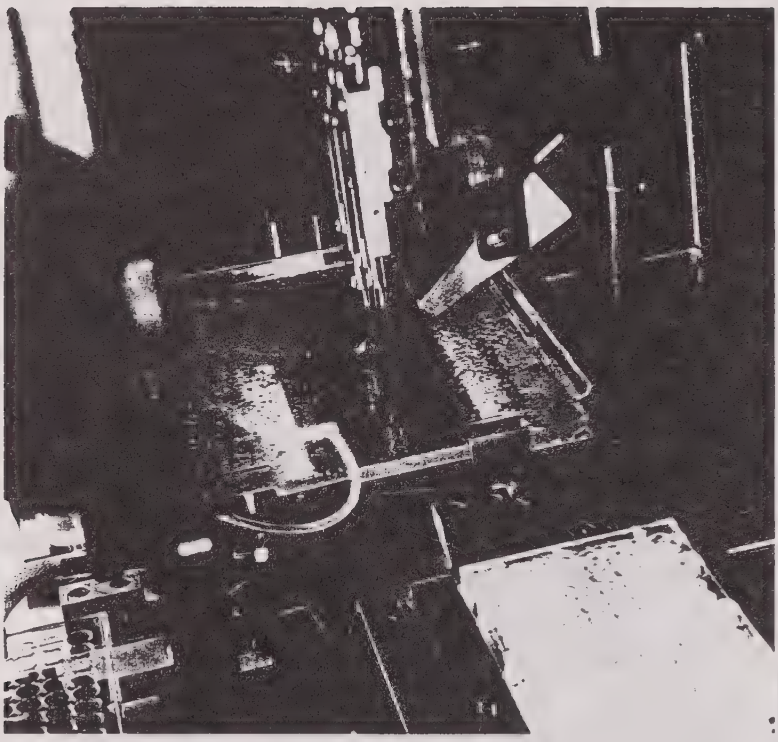
At the same time, we attempt to deduce the likely function of the protein corresponding to the partial sequence. Once we have predicted the protein's structure, we classify it according to its similarity to the structures of known proteins. Sometimes we find a match with another human protein, but often we notice a match with one from a bacterium, fungus, plant or insect: other organisms produce many proteins similar in function to those of humans. Our computers continually update these provisional classifications.

Three years ago, for example, we predicted that genes containing four specific contigs would each produce proteins similar to those known to correct mutations in the DNA of bacteria and yeast. Because researchers had learned that failure to repair mutations can cause colon cancer, we started to work out the full sequences of the four genes. When a prominent colon cancer researcher later approached us for help in identifying genes that might cause that illness—he already knew about one such gene—we were able to tell him that we were already working with three additional genes that might be involved.

Subsequent research has confirmed that mutations in any one of the four genes can cause life-threatening colon, ovarian or endometrial cancer. As many as one in every 200 people in North America and Europe carry a mutation in one of these mismatch repair genes, as they are called. Knowing this, scientists can develop tests to assess the mismatch repair genes in people who have relatives with these cancers. If the people who are tested display a genetic predisposition to illness, they can be monitored closely. Prompt detection of tumors can lead to lifesaving surgery, and such tests have already been used in clinical research to identify people at risk.

Our database now contains more than a million cDNA-derived partial gene sequences, sorted into 170,000 contigs. We think we have partial sequences from almost all expressed human genes. One indication is that when other scientists log gene sequences into public databases, we find that we already have a partial sequence for more than 95 percent of them. Piecing together partial sequences frequently uncovers entire new genes. Overall more than half of the new genes we identify have a resemblance to known genes that have been assigned a probable function. As time goes by, this proportion is likely to increase.

If a tissue gives rise to an unusually large number of cDNA sequences that derive from the same gene, it provides an indication that the gene in question is producing copious amounts of mRNA. That generally happens when the cells are producing large amounts of the cor-



ROBOT used to distinguish bacterial colonies that have picked up human DNA sequences is at the top. The instrument's arms ignore colonies that are blue, the sign that they contain no human DNA. By analyzing the sequences in the bacteria, researchers can identify human genes.

BERND AUERS

responding protein, suggesting that the protein may be doing a particularly vital job. HGS also pays particular attention to genes that are expressed only in a narrow range of tissues, because such genes are most likely to be useful for intervening in diseases affecting those tissues. Of the thousands of genes we have discovered, we have identified about 300 that seem especially likely to be medically important.

New Genes, New Medicines

Using the partial cDNA sequence technique for gene discovery, researchers have for the first time been able to assess how many genes are devoted to each of the main cellular functions, such as defense, metabolism and so on. The vast store of unique information from partial cDNA sequences offers new possibilities for medical science. These opportunities are now being systematically explored.

Databases such as ours have already proved their value for finding proteins that are useful as signposts of disease. Prostate cancer is one example. A widely used test for detecting prostate cancer measures levels in the blood of a protein called prostate specific antigen. Patients who have prostate cancer often exhibit unusually high levels. Unfortunately, slow-growing, relatively benign tumors as well as malignant tumors requiring aggressive therapy can cause elevated levels of the antigen, and so the test is ambiguous.

HGS and its partners have analyzed mRNAs from multiple samples of healthy prostate tissue as well as from benign and malignant prostate tumors. We found about 300 genes that are expressed in the prostate but in no other tissue; of these, about 100 are active only in prostate tumors, and about 20 are expressed only in tumors rated by pathologists as malignant. We and our commercial partners are using these 20 genes and their protein products to devise tests to identify malignant prostate disease. We have similar work under way for breast, lung, liver and brain cancers.

Databases of partial cDNA sequences can also help find genes responsible for rare diseases. Researchers have long known, for example, that a certain form of blindness in children is the result of an inherited defect in the chemical breakdown of the sugar galactose. A search of our database revealed two previously unknown human genes whose corresponding proteins were predicted to be structurally similar to known galactose-metabolizing enzymes in yeast and bacteria. Investigators quickly confirmed that inherited defects in either of these two genes cause this type of blindness. In the future, the enzymes or the genes themselves might be used to prevent the affliction.

Partial cDNA sequences are also establishing an impressive record for helping researchers to find smaller molecules that are candidates to be new treatments. Methods for creating and testing small-molecule drugs—the most common type—have improved dramatically in the past few years. Automated equipment can rapidly screen natural and synthetic compounds for their ability to affect a human protein involved in disease, but the limited number of known protein targets has delayed progress. As more human proteins are investigated, progress should accelerate. Our work is now providing more than half of SmithKline Beecham's leads for potential products.

Databases such as ours not only make it easier to screen molecules randomly for useful activity. Knowing a protein's structure enables scientists to custom-design drugs to interact in a specific way with the protein. This technique, known as rational drug design, was used to create some of the new protease inhibitors that are proving effective against HIV (although our database was not involved in this particular effort). We are confident that partial cDNA sequences will allow pharmacologists to make more use of rational drug design.

One example of how our database has already proved useful concerns cells known as osteoclasts, which are normally present in bone; these cells produce an enzyme capable of degrading bone tissue. The enzyme appears to be produced in excess in some disease states, such as osteoarthritis and osteoporosis. We found in our computers a sequence for a gene expressed in osteoclasts that appeared to code for the destructive enzyme; its sequence was similar to that of a gene known to give rise to an enzyme that degrades cartilage. We confirmed that the osteoclast gene was responsible for the degradative enzyme and also showed that it is not expressed in other tissues. Those discoveries meant we could invent ways to thwart the gene's protein without worrying that the methods would harm other tissues. We then made the protein, and SmithKline Beecham has used it to identify possible therapies by a combination of high-throughput screening and rational drug design. The company has also used our database to screen for molecules that might be used to treat atherosclerosis.

Protein	Activity	Possible [illegible]
Keratinocyte growth factor	Stimulates regrowth of skin	Healing wounds , stimulating hair growth, protecting against chemotherapy's side effects
Myeloid progenitor Inhibitory protein 1	**Prevents chemotherapy drugs from killing bone marrow cells**	**Protecting against chemotherapy's side effects**
Motor neuron growth factor	Prevents trauma-induced motor neuron death	Treating Lou Gehrig's disease, traumatic nerve injury, stroke and muscle atrophy in aging
Monocyte colony inhibitory factor	**Inhibits macrophages**	**Treating rheumatoid arthritis and other autoimmune and macrophage-related diseases**

JENNIFER C. CHRISTIANSEN

HUMAN PROTEINS made after their genes were discovered at Human Genome Sciences include several that demonstrate powerful effects in isolated cells and in experimental animals. These examples are among a number of human proteins now being tested to discover their possible medical value.

One extremely rich lode of genes and proteins, from a medical point of view, is a class known as G-protein coupled receptors. These proteins span the cell's outer membrane and convey biological signals from other cells into the cell's interior. It is likely that drugs able to inhibit such vital receptors could be used to treat diseases as diverse as hypertension, ulcers, migraine, asthma, the common cold and psychiatric disorders. HGS has found more than 70 new G-protein coupled receptors. We are now testing their effects by introducing receptor genes we have discovered into cells and evaluating how the cells that make the encoded proteins respond to various stimuli. Two genes that are of special interest produce proteins that seem to be critically involved in hypertension and in adult-onset diabetes. Our partners in the pharmaceutical industry are searching for small molecules that should inhibit the biological signals transmitted by these receptors.

Last but not least, our research supports our belief that some of the human genes and proteins we are now discovering will, perhaps in modified form, themselves constitute new therapies. Many human proteins are already used as drugs; insulin and clotting factor for hemophiliacs are well-known examples. Proteins that stimulate the production of blood cells are also used to speed patients' recovery from chemotherapy.

The proteins of some 200 of the full-length gene sequences HGS has uncovered have possible applications as medicines. We have made most of these proteins and have instituted tests of their activity on cells. Some of them are also proving promising in tests using experimental animals. The proteins include several chemokines, molecules that stimulate immune system cells.

Developing pharmaceuticals will never be a quick process, because medicines, whether proteins, genes or small molecules, have to be extensively tested. Nevertheless, partial cDNA sequences can speed the discovery of candidate therapies. HGS allows academic researchers access to much of its database, although we ask for an agreement to share royalties from any ensuing products.

The systematic use of automated and computerized methods of gene discovery has yielded, for the first time, a comprehensive picture of where different genes are expressed—the anatomy of human gene expression. In addition, we are starting to learn about the changes in gene expression in disease. It is too early to know exactly when physicians will first successfully use this knowledge to treat disease. Our analyses predict, however, that a number of the resulting therapies will form mainstays of 21st-century medicine. ◼

To obtain high-quality reprints of this article, please see page 123.

The Author

WILLIAM A. HASELTINE is chairman of the board of directors and chief executive officer of Human Genome Sciences in Rockville, Md. He has a doctorate in biophysics from Harvard University and from 1976 to 1993 was a professor with appointments at the Dana-Farber Cancer Institute, Harvard Medical School and Harvard School of Public Health. Haseltine has won numerous scientific awards for his research on cancer and AIDS; he has also been awarded more than 50 patents. Since 1981 he has founded seven biotechnology companies. In 1988, together with Flossie Wong-Staal, Haseltine wrote in *Scientific American* about the molecular biology of the AIDS virus.

Further Reading

DEALING WITH GENES: THE LANGUAGE OF HEREDITY. Paul Berg and Maxine Singer. University Science Books, Blackwell Scientific Publications, 1992.

MUTATION OF A MUTL HOMOLOG IN HEREDITARY COLON CANCER. N. Papadopoulos in *Science*, Vol. 263, pages 1625–1629; March 18, 1994.

MUTATIONS OF TWO PMS HOMOLOGUES IN HEREDITARY NONPOLYPOSIS COLON CANCER. N. Nicolaides et al. in *Nature*, Vol. 317, pages 75–80; September 1, 1994.

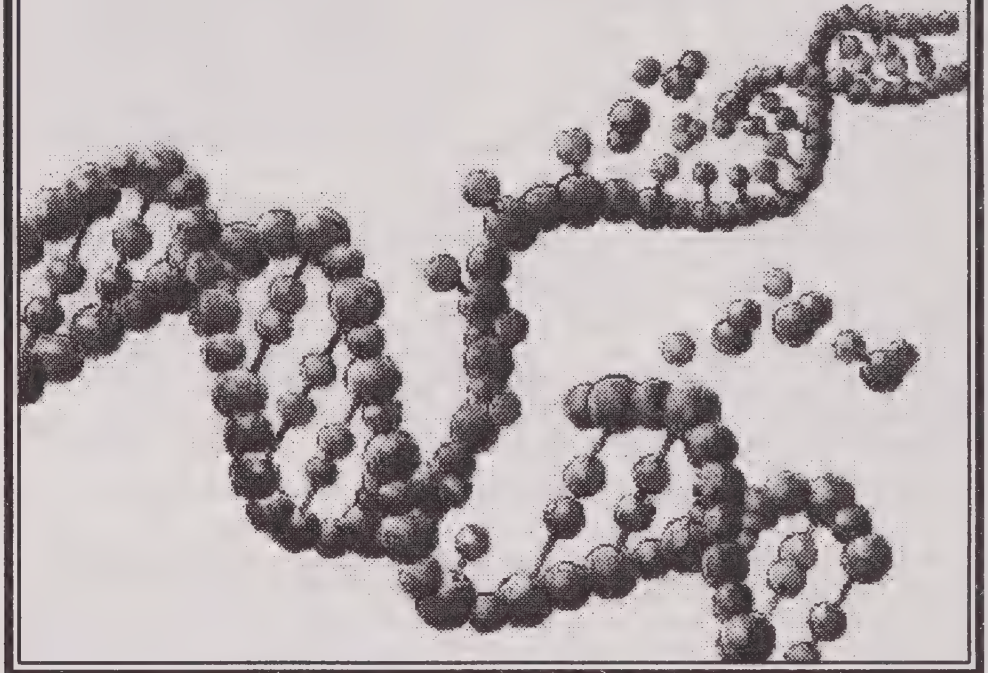
INITIAL ASSESSMENT OF HUMAN GENE DIVERSITY AND EXPRESSION PATTERNS BASED UPON 83 MILLION NUCLEOTIDES OF cDNA SEQUENCE. Mark D. Adams et al. in *Nature*, Vol. 377, Supplement, pages 3–174; September 28, 1995.

A cDNA ENCODING THE CALCITONIN GENE-RELATED PEPTIDE TYPE 1 RECEPTOR. Nambi Aiyar et al. in *Journal of Biological Chemistry*, Vol. 271, No. 19, pages 11325–11329; May 10, 1996.

CATHEPSIN K, BUT NOT CATHEPSINS B, L, OR S, IS ABUNDANTLY EXPRESSED IN HUMAN OSTEOCLASTS. Fred H. Drake et al. in *Journal of Biological Chemistry*, Vol. 271, No. 21, pages 12511–12516; May 24, 1996.

To Know
Ourselves

THE U.S. DEPARTMENT OF ENERGY
AND
THE HUMAN GENOME PROJECT



JULY 1996

Contents

FOREWORD 2

THE GENOME PROJECT—WHY THE DOE? 4
A bold but logical step

INTRODUCING THE HUMAN GENOME 6
The recipe for life

- Some definitions 6
- A plan of action 8

EXPLORING THE GENOMIC LANDSCAPE 10
Mapping the terrain

- Two giant steps: Chromosomes 16 and 19 12
- Getting down to details: Sequencing the genome 16
- Shotguns and transposons 20
- How good is good enough? 26

Sidebar: Tools of the Trade 17
Sidebar: The Mighty Mouse 24

BEYOND BIOLOGY 27
Instrumentation and informatics

- Smaller is better—And other developments 27
- Dealing with the data 30

ETHICAL, LEGAL, AND SOCIAL IMPLICATIONS 32
An essential dimension of genome research

Foreword

At the end of the road in Little Cottonwood Canyon, near Salt Lake City, Alta is a place of near-mythic renown among skiers. In time it may well assume similar status among molecular geneticists. In December 1984, a conference there, co-sponsored by the U.S. Department of Energy, pondered a single question: Does modern DNA research offer a way of detecting tiny genetic mutations—and, in particular, of observing any increase in the mutation rate among the survivors of the Hiroshima and Nagasaki bombings and their descendants? In short the answer was, Not yet. But in an atmosphere of rare intellectual fertility, the seeds were sown for a project that would make such detection possible in the future—the Human Genome Project.

In the months that followed, much deliberation and debate ensued. But in 1986, the DOE took a bold and unilateral step by announcing its Human Genome Initiative, convinced that its mission would be well served by a comprehensive picture of the human genome. The immediate response was considerable skepticism—skepticism about the scientific community's technological wherewithal for sequencing the genome at a reasonable cost and about the value of the result, even if it could be obtained economically.

Things have changed. Today, a decade later, a worldwide effort is under way to develop and apply the technologies needed to completely map and sequence the human genome, as well as the genomes of several model organisms. Technological progress has been rapid, and it is now generally agreed that this international project will produce the complete sequence of the human genome by the year 2005.

And what is more important, the value of the project also appears beyond doubt. Genome research is revolutionizing biology and biotechnology, and providing a vital thrust to the increasingly broad scope of the biological sciences. The impact that will be felt in medicine and health care alone, once we identify all human genes, is inestimable. The project has already stimulated significant investment by large corporations and prompted the creation of new companies hoping to capitalize on its profound implications.

But the DOE's early, catalytic decision deserves further comment. The organizers of the DOE's genome initiative recognized that the information the project would generate—both technological and genetic—would contribute not only to a new understanding of human biology, but also to a host of practical applications in the biotechnology industry and in the arenas of agriculture and environmental protection. A 1987 report by a DOE advisory committee provided some examples. The committee foresaw that the project could ultimately lead to the efficient production of biomass for fuel, to improvements in the resistence of plants to environmental stress, and to the practical use of genetically engineered microbes to neutralize toxic wastes. The Department thus saw far more to the genome project than a promised tool for assessing mutation rates. For example, understanding the human genome will have an enormous impact on our ability to assess, individual by individual, the risk posed by environmental exposures to toxic agents. We know that genetic differences make some of us more susceptible, and others more resistant, to such agents. Far more work must be done before we understand the genetic basis of such variability, but this knowledge will directly address the DOE's long-term mission to understand the effects of low-level exposures to radiation and other energy-related agents—especially the effects of such exposure on cancer risk. And the genome project is a long stride toward such knowledge.

The Human Genome Project has other implications for the DOE as well. In 1994, taking advantage of new capabilities developed by the genome project, the DOE formulated the Microbial Genome Initiative to sequence the genomes of bacteria of likely interest in the areas of energy production and use, environmental remediation and waste reduction, and industrial processing. As a result of this initiative, we already have complete sequences for two microbes that live under extreme conditions of temperature and pressure. Structural studies are under way to learn what is unique about the proteins of these organisms—the aim being ultimately to engineer these microbes and their enzymes for such practical purposes as waste control and environmental cleanup. (DOE-funded genetic engineering of a thermostable DNA polymerase has already produced an enzyme that has captured a large share of the several-hundred-million-dollar DNA polymerase market.)

And other little-studied microbes hint at even more intriguing possibilities. For instance, *Deinococcus radiodurans* is a species that prospers even when exposed to huge doses of ionizing radiation. This microbe has an amazing ability to repair radiation-induced damage to its DNA. Its genome is currently being sequenced with DOE support, with the hope of understanding and ultimately taking practical advantage of its unusual capabilities. For example, it might be possible to insert foreign DNA into this microbe that allows it to digest toxic organic components found in highly radioactive waste, thus simplifying the task of further cleanup. Another approach might be to introduce metal-binding proteins onto the microbe's surface that would scavenge highly radioactive isotopes out of solution.

Biotechnology, fueled in part by insights reaped from the genome project, will also play a significant role in improving the use of fossil-based resources. Increased energy demands, projected over the next 50 years, require strategies to circumvent the many problems associated with today's dominant energy systems. Biotechnology promises to help address these needs by upgrading the fuel value of our current energy resources and by providing new means for the bioconversion of raw materials to refined products—not to mention offering the possibility of entirely new biomass-based energy sources.

We have thus seen only the dawn of a biological revolution. The practical and economic applications of biology are destined for dramatic growth. Health-related biotechnology is already a multibillion-dollar success story—and is still far from reaching its potential. Other applications of biotechnology are likely to beget similar successes in the coming decades. Among these applications are several of great importance to the DOE. We can look to improvements in waste control and an exciting era of environmental bioremediation; we will see new approaches to improving energy efficiency; and we can even hope for dramatic strides toward meeting the fuel demands of the future. The insights, the technologies, and the infrastructure that are already emerging from the genome project, together with advances in fields such as computational and structural biology, are among our most important tools in addressing these national needs.

AAPatrinos

Aristides A. N. Patrinos
Director, Human Genome Project
U.S. Department of Energy

The Genome Project—Why the DOE?

A BOLD BUT LOGICAL STEP

THE BIOSCIENCES RESEARCH community is now embarked on a program whose boldness, even audacity, has prompted comparisons with such visionary efforts as the Apollo space program and the Manhattan project. That life scientists should conceive such an ambitious project is not remarkable; what is surprising—at least at first blush—is that the project should trace its roots to the Department of Energy.

For close to a half-century, the DOE and its governmental predecessors have been charged with pursuing a deeper understanding of the potential health risks posed by energy use and by energy-production technologies—with special interest focused on the effects of radiation on humans. Indeed, it is fair to say that most of what we know today about radiological health hazards stems from studies supported by these government agencies. Among these investigations are long-standing studies of the survivors of the atomic bombings of Hiroshima and Nagasaki, as well as any number of experimental studies using animals, cells in culture, and nonliving systems. Much has been learned, especially about the consequences of exposure to high doses of radiation. On the other hand, many questions remain unanswered; in particular, we have much to learn about how low doses produce their insidious effects. When present merely in low but significant amounts, toxic agents such as radiation or mutagenic chemicals work their mischief in the most subtle ways, altering only slightly the genetic instructions in our cells. The consequences can be heritable mutations too slight to produce discernible effects in a generation or two but, in their persistence and irreversibility, deeply troublesome nonetheless.

In 1980 the DOE was the first federal agency to announce an initiative to pursue a detailed understanding of the human genome.

Until recently, science offered little hope for detecting at first hand these tiny changes to the DNA that encodes our genetic program. Needed was a tool that could detect a change in one "word" of the program, among perhaps a hundred million. Then, in 1984, at a meeting convened jointly by the DOE and the International Commission for Protection Against Environmental Mutagens and Carcinogens, the question was first seriously asked: Can we, should we, sequence the human genome? That is, can we develop the technology to obtain a word-by-word copy of the entire genetic script for an "average" human being, and thus to establish a benchmark for detecting the elusive mutagenic effects of radiation and cancer-causing toxins? Answering such a question was not simple. Workshops were convened in 1985 and 1986; the issue was studied by a DOE advisory group, by the Congressional Office of Technology Assessment, and by the National Academy of Sciences; and the matter was debated publicly and privately among biologists themselves. In the end, however, a consensus emerged that we should make a start.

4

Adding impetus to the DOE's earliest interest in the human genome was the Department's stewardship of the national laboratories, with their demonstrated ability to conduct large multidisciplinary projects—just the sort of effort that would be needed to develop and implement the technological know-how needed for the Human Genome Project. Biological research programs already in place at the national labs benefited from the contributions of engineers, physicists, chemists, computer scientists, and mathematicians, working together in teams. Thus, with the infrastructure in place and with a particular interest in the ultimate results, the Department of Energy, in 1986, was the first federal agency to announce and to fund an initiative to pursue a detailed understanding of the human genome.

Of course, interest was not restricted to the DOE. Workshops had also been sponsored by the National Institutes of Health, the Cold Spring Harbor Laboratory, and the Howard Hughes Medical Institute. In 1988 the NIH joined in the pursuit, and in the fall of that year, the DOE and the NIH signed a memorandum of understanding that laid the foundation for a concerted interagency effort. The basis for this community-wide excitement is not hard to comprehend. The first impulse behind the DOE's commitment was only one of many reasons for coveting a deeper insight into the human genetic script. Defective genes directly account for an estimated 4000 hereditary human diseases—maladies such as Huntington disease and cystic fibrosis. In some such cases, a single misplaced letter among three billion can have lethal consequences. For most of us, though, even greater interest focuses on the far more common ailments in which altered genes influence but do not prescribe. Heart disease, many cancers, and some psychiatric disorders, for example, can emerge from complicated interplays of environmental factors and genetic misinformation.

The first steps in the Human Genome Project are to develop the needed technologies, then to "map" and "sequence" the genome. But in a sense, these well-publicized efforts aim only to provide the raw material for the next, longer strides. The ultimate goal is to exploit those resources for a truly profound molecular-level understanding of how we develop from embryo to adult, what makes us work, and what causes things to go wrong. The benefits to be reaped stretch the imagination. In the offing is a new era of molecular medicine characterized not by treating symptoms, but rather by looking to the deepest causes of disease. Rapid and more accurate diagnostic tests will make possible earlier treatment for countless maladies. Even more promising, insights into genetic susceptibilities to disease and to environmental insults, coupled with preventive therapies, will thwart some diseases altogether. New, highly targeted pharmaceuticals, not just for heritable diseases, but for communicable ailments as well, will attack diseases at their molecular foundations. And even gene therapy will become possible, in some cases actually "fixing" genetic errors. All of this in addition to a new intellectual perspective on who we are and where we came from.

The Department of Energy is proud to be playing a central role in propelling us toward these noble goals.

Introducing the Human Genome

THE RECIPE FOR LIFE

FOR ALL THE DIVERSITY of the world's five and a half billion people, full of creativity and contradictions, the machinery of every human mind and body is built and run with fewer than 100,000 kinds of protein molecules. And for each of these proteins, we can imagine a single corresponding gene (though there is sometimes some redundancy) whose job it is to ensure an adequate and timely supply. In a material sense, then, all of the subtlety of our species, all of our art and science, is ultimately accounted for by a surprisingly small set of discrete genetic instructions. More surprising still, the differences between two unrelated individuals, between the man next door and Mozart, may reflect a mere handful of differences in their genomic recipes—perhaps one altered word in five hundred. We are far more alike than we are different. At the same time, there is room for near-infinite variety.

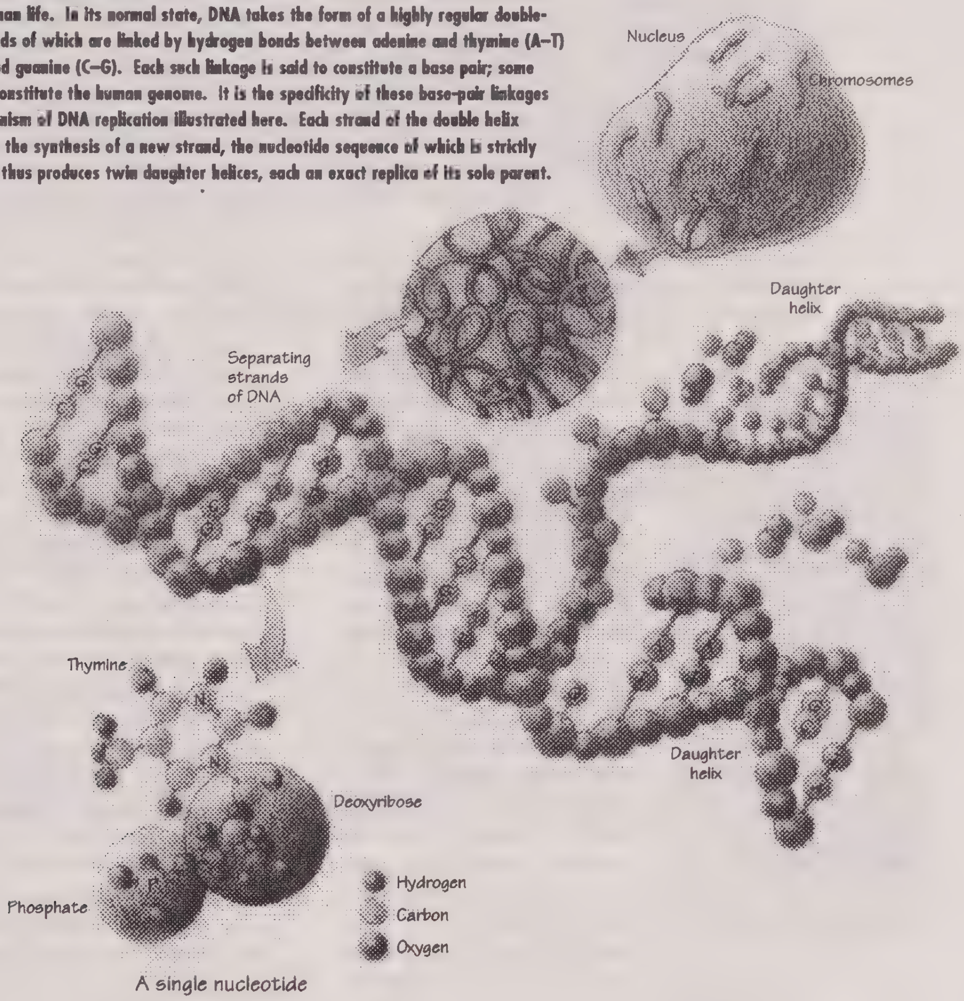
It is no overstatement to say that to decode our 100,000 genes in some fundamental way would be an epochal step toward unraveling the manifold mysteries of life.

SOME DEFINITIONS

The *human genome* is the full complement of genetic material in a human cell. (Despite five and a half billion variations on a theme, the differences from one genome to the next are minute; hence, we hear about *the* human genome—as if there were only one.) The genome, in turn, is distributed among 23 sets of *chromosomes*, which, in each of us, have been replicated and re-replicated since the fusion of sperm and egg that marked our conception. The source of our personal uniqueness, our full genome, is therefore preserved in each of our body's several trillion cells. At a more basic level, the genome is DNA, deoxyribonucleic acid, a natural polymer built up of repeating *nucleotides*, each consisting of a simple sugar, a phosphate group, and one of four nitrogenous bases. The hierarchy of structure from chromosome to nucleotide is shown in Figure 1. In the chromosomes, two DNA strands are twisted together into an entwined spiral—the famous double helix—held together by weak bonds between complementary bases, adenine (A) in one strand to thymine (T) in the other, and cytosine to guanine (C–G). In the language of molecular genetics, each of these linkages constitutes a *base pair*. All told, if we count only one of each pair of chromosomes, the human genome comprises about three billion base pairs

The specificity of these base-pair linkages underlies all that is wonderful about DNA. First, replication becomes straightforward. Unzipping the double helix provides unambiguous templates for the synthesis of daughter molecules: One helix begets two with near-perfect fidelity. Second, by a similar template-based process, depicted in Figure 2, a means is also available for producing a DNA-like messenger to the cell cytoplasm. There, this *messenger RNA*, the faithful complement of a particular DNA segment, directs the synthesis of a particular protein. Many subtleties are entailed in the synthesis of proteins, but in a schematic sense, the process is elegantly simple.

FIGURE 1. SOME DNA DETAILS. Apart from reproductive gametes, each cell of the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA. Every strand of the DNA is a huge natural polymer of repeating nucleotide units, each of which comprises a phosphate group, a sugar (deoxyribose), and a base (either adenine, thymine, cytosine, or guanine). Every strand thus embodies a code of four characters (A's, T's, C's, and G's), the recipe for the machinery of human life. In its normal state, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between adenine and thymine (A–T) and between cytosine and guanine (C–G). Each such linkage is said to constitute a base pair; some three billion base pairs constitute the human genome. It is the specificity of these base-pair linkages that underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand, the nucleotide sequence of which is strictly determined. Replication thus produces twin daughter helices, each an exact replica of its sole parent.



Every *protein* is made up of one or more polypeptide chains, each a series of (typically) several hundred molecules known as *amino acids*, linked by so-called peptide bonds. Remarkably, only 20 different kinds of amino acids suffice as the building blocks for all human proteins. The synthesis of a protein chain, then, is simply a matter of specifying a particular sequence of amino acids. This is the role of the messenger RNA. (The same nitrogenous bases are at work in RNA as in DNA, except that uracil takes the place of the DNA base thymine.) Each linear sequence of three bases (both in RNA and in DNA) corresponds uniquely to a single amino acid. The RNA sequence AAU thus dictates that the amino acid asparagine should be added to a polypeptide chain, GCA specifies alanine—and so on. A segment of the chromosomal DNA that directs the synthesis of a single type of protein constitutes a single *gene*.

A PLAN OF ACTION

In 1990 the Department of Energy and the National Institutes of Health developed a joint research plan for their genome programs, outlining specific goals for the ensuing five years. Three years later, emboldened by progress that was on track or even ahead of schedule, the two agencies put forth an updated five-year plan. Improvements in technology, together with the experience of three years, allowed an even more ambitious prospect.

In broad terms, the revised plan includes goals for genetic and physical mapping of the genome, DNA sequencing, identifying and locating genes, and pursuing further developments in technology and informatics. To a large extent, the following pages are devoted to a discussion of just what these goals mean, and what part the DOE is playing in pursuing them. In addition, the plan emphasizes the continuing importance of the ethical, legal, and social implications of genome research, and it underscores the critical roles of scientific training, technology transfer, and public access to research data and materials. Most of the goals focus on the human genome, but the importance of continuing research on widely studied "model organisms" is also explicitly recognized.

The plan includes goals for genetic and physical mapping, DNA sequencing, identifying and locating genes, and pursuing further developments in technology and informatics.

Among the scientific goals of human genome research, several are especially notable, as they provide clear milestones for future progress. In reciting them, however, it is important to note an underlying assumption of adequate research support. Such support is obviously crucial if the joint plan is to succeed. Some of the central goals for 1993–98 follow:

- Complete a genetic linkage map at a resolution of two to five centimorgans by 1995—As discussed on page 10, this goal was far surpassed by the fall of 1994.
- Complete a physical map at a resolution of 100 kilobases by 1998—This implies a genome map with 30,000 "signposts," separated by an average of 100,000 base pairs. Further, each signpost will be a *sequence-tagged site*, a stretch of DNA with a unique and well-defined DNA sequence. Such a map will greatly facilitate "production sequencing" of the entire genome. By the end of 1995, molecular biologists were halfway to this goal: A physical map was announced with 15,000 sequence-tagged signposts. Physical mapping is discussed on pages 10–16.
- By 1998 develop the capacity to sequence 50 million base pairs per year in long continuous segments—Adequate fiscal investment and continuing progress beyond 1998 should then produce a fully sequenced human genome by the year 2005 or earlier. Sequencing is the subject of pages 16–26.
- Develop efficient methods for identifying and locating known genes on physical maps or sequenced DNA—The goals here are less quantifiable, but the aim is central to the Human Genome Project: to home in on and ultimately to understand the most important human genes, namely, the ones responsible for serious diseases and those crucial for healthy development and normal functions.
- Pursue technological developments in areas such as automation and robotics—A continuing emphasis on technological advance is critical. Innovative technologies, such as those described on pages 27–30, are the necessary underpinnings of future large-scale sequencing efforts.
- Continue the development of database tools and software for managing and interpreting genome data—This is the area of informatics, discussed on pages 30–31. The challenge is not so much the volume of data, but rather the need to

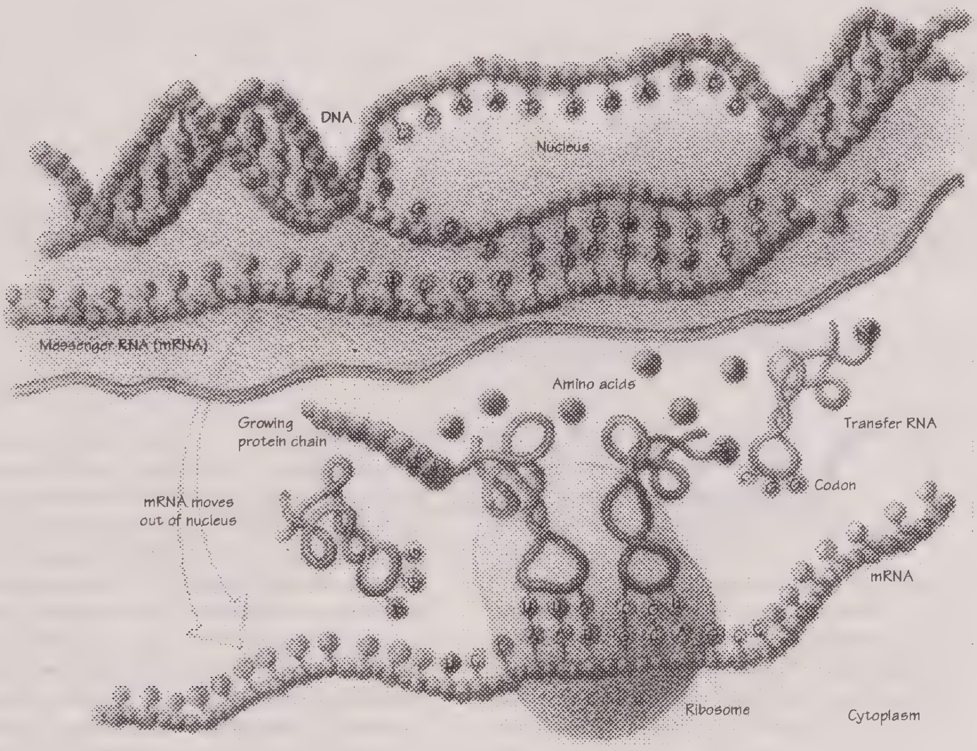


FIGURE 2. FROM GENES TO PROTEINS. In the cell nucleus, RNA is produced by transcription, in much the same way that DNA replicates itself. RNA, however, substitutes the sugar ribose for deoxyribose and the base uracil for thymine, and is usually single-stranded. One form of RNA, messenger RNA or mRNA, conveys the DNA recipe for protein synthesis to the cell cytoplasm. There, bound temporarily to a cytoplasmic particle known as a ribosome, each three-base codon of the mRNA links to a specific form of transfer RNA (tRNA) containing the complementary three-base sequence. This tRNA, in turn, transfers a single amino acid to a growing protein chain. Each codon thus unambiguously directs the addition of one amino acid to the protein. On the other hand, the same amino acid can be added by different codons; in this illustration, the mRNA sequences GCA and GCC are both specifying the addition of the amino acid alanine (Ala).

mount a system compatible with researchers around the world, and one that will allow scientists to contribute new data and to freely interrogate the existing databases. The ultimate measure of success will be the ease with which biologists can fruitfully use the information produced by the genome project.

- Continue to explore the ethical, legal, and social implications of genome research — Much emphasis continues to be placed on issues of privacy and the fair use of genetic information. New goals focus on defining additional pertinent issues and developing policy responses to them, disseminating policy options regarding genetic testing services, fostering greater acceptance of human genetic variation, and enhancing public and professional education that is sensitive to sociocultural and psychological issues. This side of the genome project is discussed on pages 32–33.

Exploring the Genomic Landscape

MAPPING THE TERRAIN

One of the central goals of the Human Genome Project is to produce a detailed "map" of the human genome. But, just as there are topographic maps and political maps and highway maps of the United States, so there are different kinds of genome maps, the variety of which is suggested in Figure 3. One type, a *genetic linkage map*, is based on careful analyses of human inheritance patterns. It indicates for each chromosome the whereabouts of genes or other "heritable markers," with distances measured in centimorgans, a measure of recombination frequency. During the formation of sperm and egg cells, a process of genetic recombination—or "crossing over"—occurs in which pieces of genetic material are swapped between paired chromosomes. This process of chromosomal scrambling accounts for the differences invariably seen even in siblings (apart from identical twins). Logically, the closer two genes are to each other on a single chromosome, the less likely they are to get split up during genetic recombination. When they are close enough that the chances of being separated are only one in a hundred, they are said to be separated by a distance of one centimorgan.

> *Just as there are topographic maps and political maps and highway maps, so there are different kinds of genome maps.*

The role of human pedigrees now becomes clear. By studying family trees and tracing the inheritance of diseases and physical traits, or even unique segments of DNA identifiable only in the laboratory, geneticists can begin to pin down the relative positions of these genetic markers. By the end of 1994, a comprehensive map was available that included more than 5800 such markers, including genes implicated in cystic fibrosis, myotonic dystrophy, Huntington disease, Tay-Sachs disease, several cancers, and many other maladies. The average gap between markers was about 0.7 centimorgan.

Other maps are known as *physical maps*, so called because the distances between features are measured not in genetic terms, but in "real" physical units, typically, numbers of base pairs. A close analogy can thus be drawn between physical maps and the road maps familiar to us all. Indeed, the analogy can be extended further. Just as small-scale road maps may show only large cities and indicate distances only between major features, so a low-resolution physical map includes only a relative sprinkling of chromosomal landmarks. A well-known low-resolution physical map, for example, is the familiar chromosomal map, showing the distinctive staining patterns that can be seen in the light microscope. Further, by a process known as *in situ hybridization*, specific segments of DNA can be targeted in intact chromosomes by using complementary strands synthesized in the laboratory. These laboratory-made "probes" carry a fluorescent or radioactive

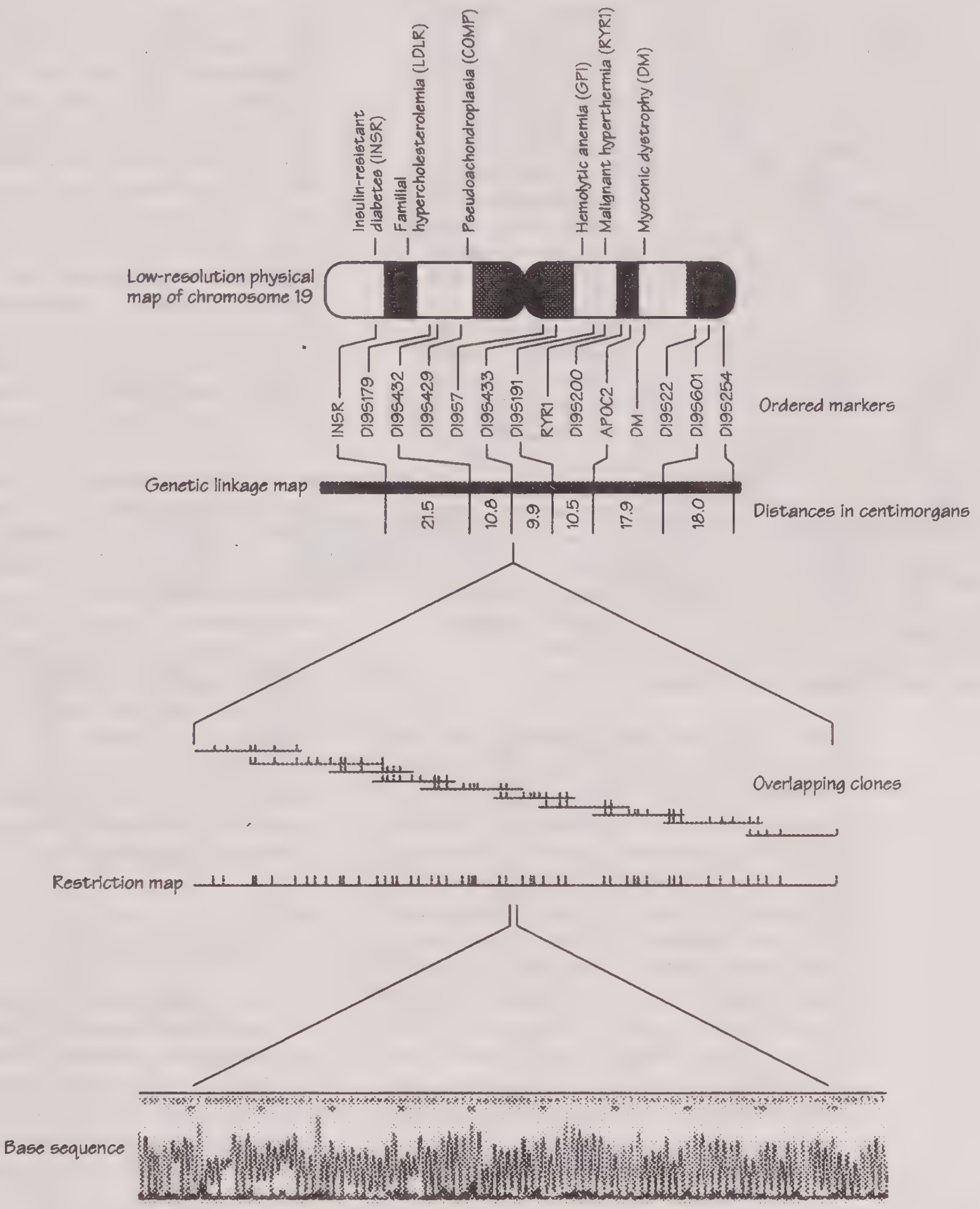

Insulin-resistant diabetes (INSR)
Familial hypercholesterolemia (LDLR)
Pseudoachondroplasia (COMP)
Hemolytic anemia (GPI)
Malignant hyperthermia (KYR1)
Myotonic dystrophy (DM)
Low-resolution physical map of chromosome 19
INSR
D19S179
D19S432
D19S429
D19S7
D19S433
D19S191
KYR1
D19S200
APOC2
DM
D19S22
D19S601
D19S254
Ordered markers
Genetic linkage map
21.5
10.8
9.9
10.5
17.9
18.0
Distances in centimorgans
Overlapping clones
Restriction map
Base sequence

FIGURE 3. GENOMIC GEOGRAPHY. The human genome can be mapped in a number of ways. The familiar and reproducible banding pattern of the chromosomes constitutes one kind of physical map, and in many cases, the positions of genes or other heritable markers have been localized to one band or another. More useful are genetic linkage maps, on which the relative positions of markers have been established by studying how frequently the markers are separated during a natural process of chromosomal shuffling called genetic recombination. The cryptically coded ordered markers near the top of this figure are physically mapped to specific regions of chromosome 19; some of them also constitute a low-resolution genetic linkage map. (Hundreds of genes and other markers have been mapped on chromosome 19; only a few are indicated here. See Figure 5 for a display of mapped genes.) A higher-resolution physical map might describe, as shown here, the cutting sites (the short vertical lines) for certain DNA-cleaving enzymes. The overlapping fragments that allow such a map to be constructed are then the resources for obtaining the ultimate physical map, the base-pair sequence for the human genome. At the bottom of this figure is an example of output from an automatic sequencing machine.

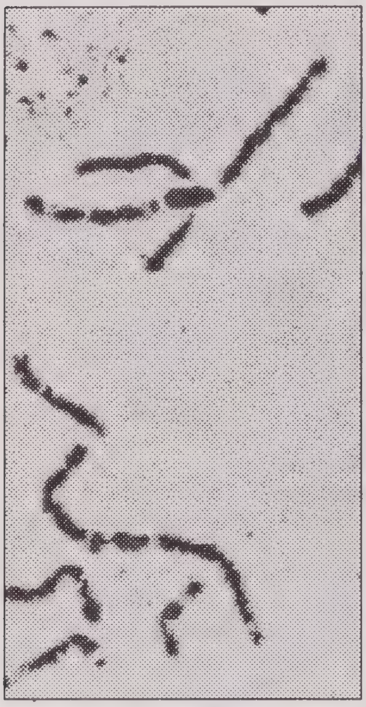


FIGURE 4. FISHING FOR GENES. Fluorescence in situ hybridization (FISH) probes are strands of DNA that have been labeled with fluorescent dye molecules. The probes bind uniquely to complementary strands of chromosomal DNA, thus pinpointing the positions of target DNA sequences. In this example, one probe, whose fluorescence signal is shown in red, binds specifically to a gene (DSRAD) that codes for an important RNA-modifying enzyme. A second probe, whose signal appears in green, binds to a marker sequence whose location was already known. The previously unknown location of the DSRAD gene was thus accurately mapped to a narrow region on the long arm of chromosome 1.

label, which can then be detected and thus pinpointed on a specific region of the chromosome. Figure 4 shows some results of fluorescence in situ hybridization (FISH). Of particular interest are probes known as cDNA (for *complementary DNA*), which are synthesized by using molecules of messenger RNA as templates. These molecules of cDNA thus hybridize to "expressed" chromosomal regions—regions that directly dictate the synthesis of proteins. However, a physical map that depended only on in situ hybridization would be a fairly coarse one. Fluorescent tags on intact chromosomes cannot be resolved into separate spots unless they are two to five million base pairs apart.

Fortunately, means are also available to produce physical maps of much higher resolution—analogous to large-scale county maps that show every village and farm road, and indicate distances at a similar level of detail. Just such a detailed physical map is one that emerges from the use of *restriction enzymes*—DNA-cleaving enzymes that serve as highly selective microscopic scalpels (see "Tools of the Trade," pages 17–19). A typical restriction enzyme known as *Eco*RI, for example, recognizes the DNA sequence GAATTC and selectively cuts the double helix at that site. One use of these handy tools involves cutting up a selected chromosome into small pieces, then cloning and ordering the resulting fragments. The *cloning*, or copying, process is a product of recombinant DNA technology, in which the natural reproductive machinery of a "host" organism—a bacterium or a yeast, for example—replicates a "parasitic" fragment of human DNA, thus producing the multiple copies needed for further study (see "Tools of the Trade"). By cloning enough such fragments, each overlapping the next and together spanning long segments (or even the entire length) of the chromosome, workers can eventually produce an ordered library of clones. Each contiguous block of ordered clones is known as a *contig* (a small one is shown in Figure 3), and the resulting map is a contig map. If a gene can be localized to a single fragment within a contig map, its physical location is thereby accurately pinned down. Further, these conveniently sized clones become resources for further studies by researchers around the world—as well as the natural starting points for systematic sequencing efforts.

TWO GIANT STEPS: CHROMOSOMES 16 AND 19

One of the signal achievements of the DOE genome effort so far is the successful physical mapping of chromosomes 16 and 19. The high-resolution chromosome 19 map, constructed at the Lawrence Livermore National Laboratory, is based on restriction fragments cloned in *cosmids*, synthetic cloning "vectors" modeled after bacteria-infecting viruses known as bacteriophages. Like a phage, a cosmid hijacks the cellular machinery of a bacterium to mass-produce its own genetic material, together with any "foreign" human DNA that has been smuggled into it. The foundation of the chromosome 19 map is a large set of cosmid contigs that were assembled by automated analysis of overlapping

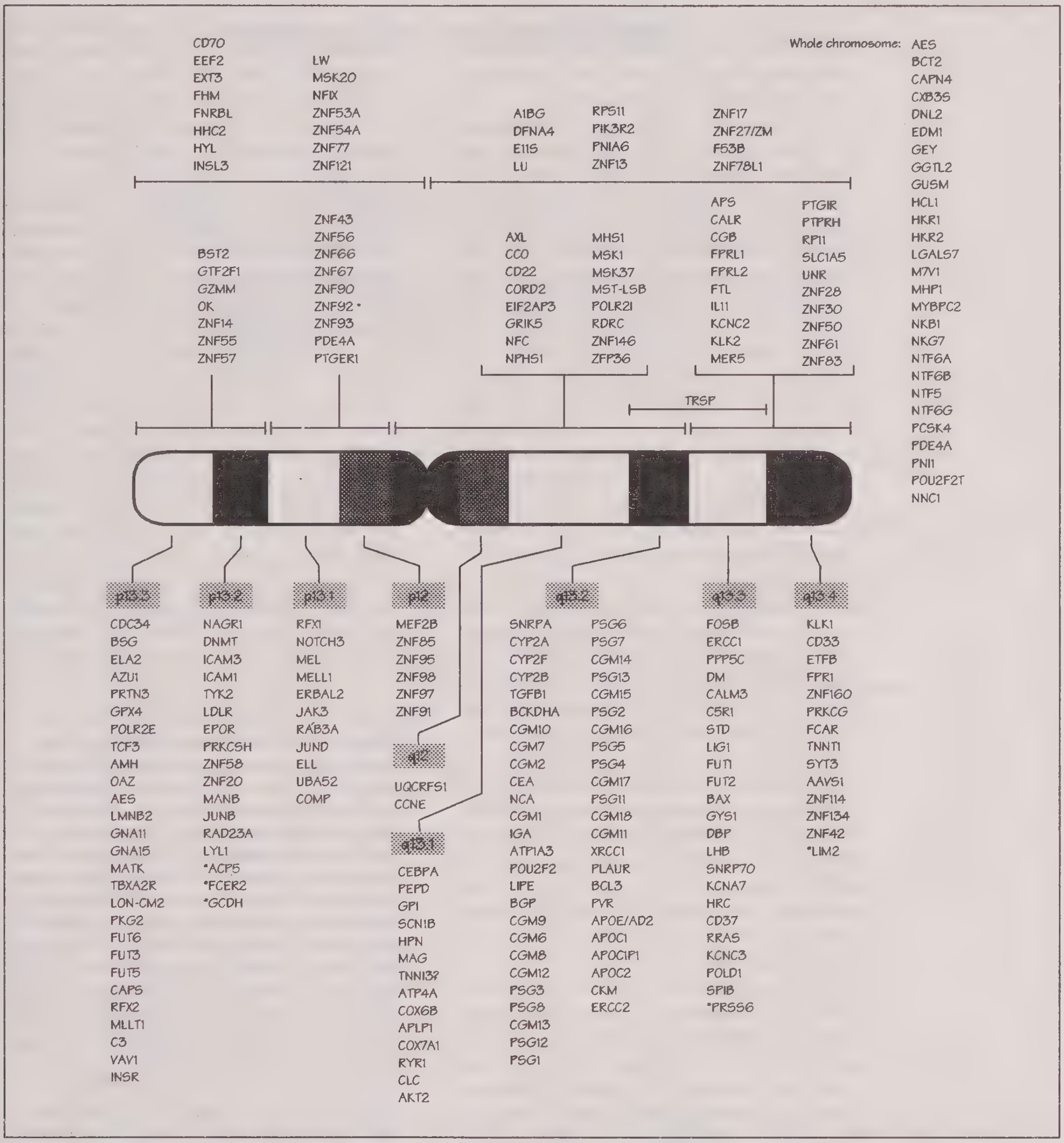

CD70
EEF2
EXT3
FHM
FNRBL
HHC2
HYL
INSL3
LW
MSK20
NFIX
ZNF53A
ZNF54A
ZNF77
ZNF121
A1BG
DFNA4
E11S
LU
RPS11
PIK3R2
PNIA6
ZNF13
ZNF17
ZNF27/ZM
F53B
ZNF78L1
Whole chromosome:
AES
BCT2
CAPN4
CXB3S
DNL2
EDM1
GEY
GGTL2
GUSM
HCL1
HKR1
HKR2
LGALS7
M7V1
MHP1
MYBPC2
NKB1
NKG7
NTF6A
NTF6B
NTF5
NTF6G
PCSK4
PDE4A
PNI1
POU2F2T
NNC1
BST2
GTF2F1
GZMM
OK
ZNF14
ZNF55
ZNF57
ZNF43
ZNF56
ZNF66
ZNF67
ZNF90
ZNF92 *
ZNF93
PDE4A
PTGER1
AXL
CCO
CD22
CORD2
EIF2AP3
GRIK5
NFC
NPHS1
MHS1
MSK1
MSK37
MST-LSB
POLR2I
RDRC
ZNF146
ZFP36
APS
CALR
CGB
FPRL1
FPRL2
FTL
IL11
KCNC2
KLK2
MER5
PTGIR
PTPRH
RP11
SLC1A5
UNR
ZNF28
ZNF30
ZNF50
ZNF61
ZNF83
TRSP
p13.3
CDC34
BSG
ELA2
AZU1
PRTN3
GPX4
POLR2E
TCF3
AMH
OAZ
AES
LMNB2
GNA11
GNA15
MATK
TBXA2R
LON-CM2
PKG2
FUT6
FUT3
FUT5
CAPS
RFX2
MLLT1
C3
VAV1
INSR
p13.2
NAGR1
DNMT
ICAM3
ICAM1
TYK2
LDLR
EPOR
PRKCSH
ZNF58
ZNF20
MANB
JUNB
RAD23A
LYL1
*ACP5
*FCER2
*GCDH
p13.1
RFX1
NOTCH3
MEL
MELL1
ERBAL2
JAK3
RAB3A
JUND
ELL
UBA52
COMP
p12
MEF2B
ZNF85
ZNF95
ZNF98
ZNF97
ZNF91
q12
UQCRFS1
CCNE
q13.1
CEBPA
PEPD
GPI
SCN1B
HPN
MAG
TNNI3?
ATP4A
COX6B
APLP1
COX7A1
RYR1
CLC
AKT2
q13.2
SNRPA
CYP2A
CYP2F
CYP2B
TGFB1
BCKDHA
CGM10
CGM7
CGM2
CEA
NCA
CGM1
IGA
ATP1A3
POU2F2
LIPE
BGP
CGM9
CGM6
CGM8
CGM12
PSG3
PSG8
CGM13
PSG12
PSG1
PSG6
PSG7
CGM14
PSG13
CGM15
PSG2
CGM16
PSG5
PSG4
CGM17
PSG11
CGM18
CGM11
XRCC1
PLAUR
BCL3
PVR
APOE/AD2
APOC1
APOC1P1
APOC2
CKM
ERCC2
q13.3
FOSB
ERCC1
PPP5C
DM
CALM3
C5R1
STD
LIG1
FUT1
FUT2
BAX
GYS1
DBP
LHB
SNRP70
KCNA7
HRC
CD37
RRAS
KCNC3
POLD1
SPIB
*PRSS6
q13.4
KLK1
CD33
ETFB
FPR1
ZNF160
PRKCG
FCAR
TNNT1
SYT3
AAVS1
ZNF114
ZNF134
ZNF42
*LIM2

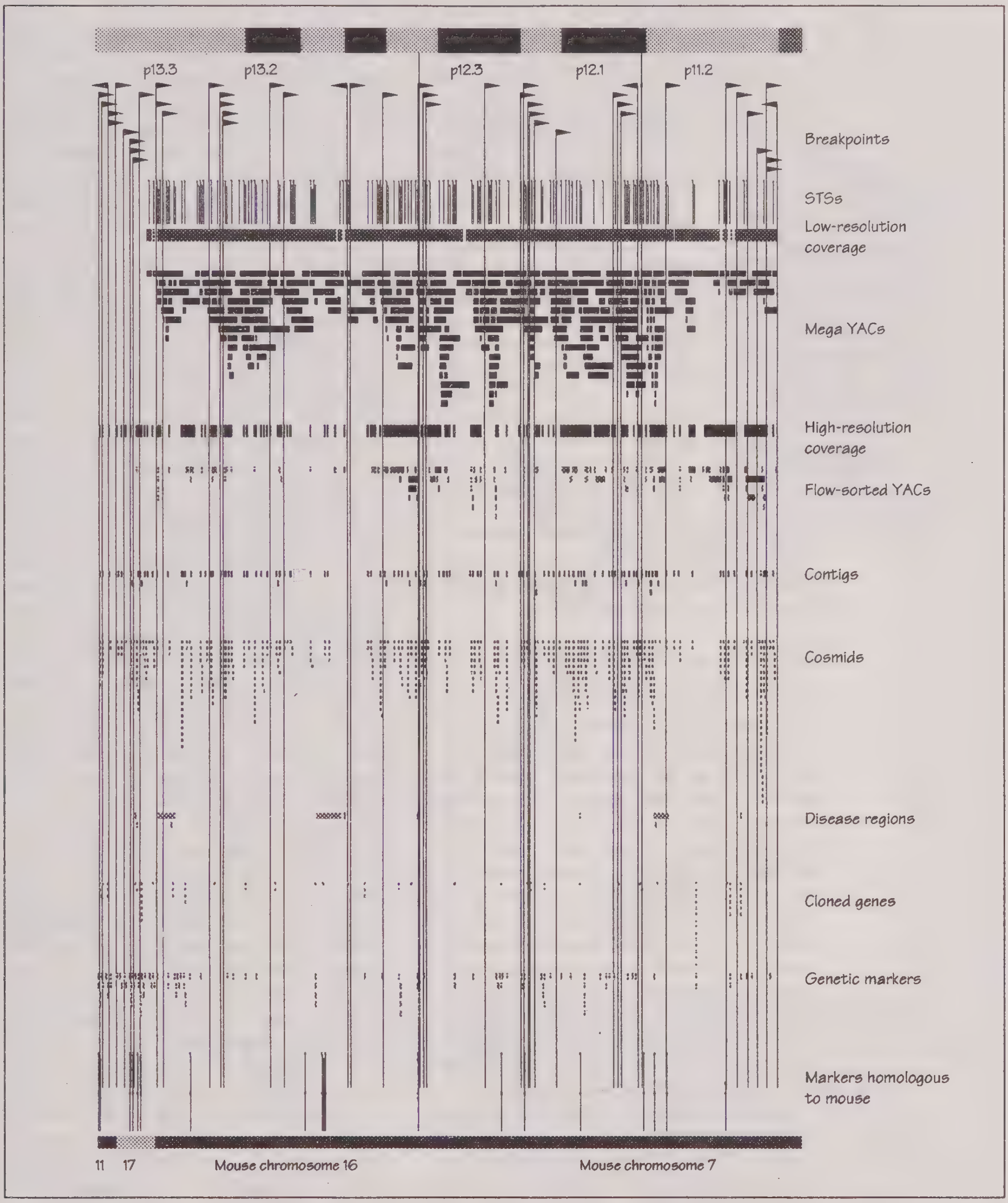
FIGURE 5. AN EMERGING GENE MAP. More than 250 genes have already been mapped to chromosome 19. Those listed on the lower half of this illustration have been assigned to specific cosmids and (except for those marked with asterisks) have been ordered on the Livermore physical map. Their positions are therefore known with far greater accuracy than shown here. The genes listed above the chromosome have been mapped to larger regions of the chromosome—or merely localized to chromosome 19 generally—and have not yet been assigned to cosmids in the Livermore database. The text mentions several of the most important genes mapped so far. Others include INSR, which codes for an insulin receptor and is involved in adult-onset diabetes; LDLR, a gene for a low-density lipoprotein receptor involved in hypercholesterolemia; and ERCC2, a DNA repair gene implicated in one form of xeroderma pigmentosum.

but unordered restriction fragments. These contigs span an estimated 54 million base pairs, more than 95 percent of the chromosome, excluding the centromere.

Most of the contigs have been mapped by fluorescence in situ hybridization to visible chromosomal bands. Further, more than 200 cosmids have been more accurately ordered along the chromosome by a high-resolution FISH technique in which the distances between cosmids are determined with a resolution of about 50,000 base pairs. This ordered FISH map, with cosmid reference points separated by an average of 230,000 base pairs, provides the essential framework to which other cosmid contigs can be anchored. Moreover, the *Eco*RI restriction sites have been mapped on more than 45 million base pairs of the overall cosmid map. Over 450 genes and genetic markers have also been localized on this map, of which nearly 300 have been incorporated into the ordered map. Figure 5 shows the locations of the mapped genes. Among these genes is the one responsible for the most common form of adult muscular dystrophy (DM), which was identified in 1992 by an international consortium that included Livermore scientists. A second important disease gene (COMP), responsible for a form of dwarfism known as pseudoachondroplasia, has also been identified. And yet another gene, one linked to a form of congenital kidney disease, has been localized to a single contig spanning one million base pairs, but has not yet been precisely pinpointed. About 2000 other genes are likely to be found eventually on chromosome 19.

In a similar effort, the Los Alamos National Laboratory Center for Human Genome Studies has completed a highly integrated map of chromosome 16, a chromosome that contains genes linked to blood disorders, a second form of kidney disease, leukemia, and breast and prostate cancers. A readable display of this integrated map covers a sheet of paper more than 15 feet long; a portion of it, much reduced and showing only some of its central features, is reproduced here as Figure 6. The framework for the Los Alamos effort is yet another kind of map, a "cytogenetic breakpoint map" based on 78 lines of cultured cells, each a hybrid that contains mouse chromosomes and a fragment of human chromosome 16. Natural breakpoints in chromosome 16 are thus identified, leading to a breakpoint map that divides the chromosome into segments whose lengths average 1.1 million base pairs. Anchored to this framework are a low-resolution contig map based on YAC clones and a high-resolution contig map based largely on cosmids (for more on YACs, yeast artificial chromosomes, see "Tools of the Trade," pages 17–19). The low-resolution map, comprising 700 YACs from a library constructed by the Centre d'Etude du Polymorphisme Humain (CEPH), provides practically complete coverage of the chromosome, except the highly repetitive DNA in the centromere region. The high-resolution map comprises some 4000 cosmid clones, assembled into about 500 contigs covering 60 percent of the chromosome. In addition, it includes 250 smaller YAC clones that have been merged with the cosmid contig map. The cosmid contig map

FIGURE 6. MAPPING CHROMOSOME 16. **This much-reduced physical map of the short arm of human chromosome 16 summarizes the progress made at Los Alamos toward a complete map of the chromosome. A legible, fully detailed map of the chromosome is more than 15 feet long; only a few features of the map can be described here. Just below the schematic chromosome, the black arrowheads and the vertical lines extending the full length of the page signify "breakpoints" and indicate the portions of the chromosome maintained in separate cell cultures. The cultured portions typically extend from a breakpoint to one end of the chromosome. These breakpoints establish the framework for the Los Alamos mapping effort. Within this framework, some 700 megaYACs (shown in black) provide low-resolution coverage for essentially the entire chromosome. Smaller flow-sorted YACs (light blue, red, and black), together with about 4000 cosmids, assembled into about 500 cosmid contigs (blue and red), establish high-resolution coverage for 60% of the chromosome. Sequence-tagged sites (STSs) are shown as colored vertical lines above the megaYACs, and genes (green) and genetic markers (pink) that have been localized only to the breakpoint map are shown near the bottom. Also shown are cloned and uncloned disease regions, as well as those markers whose analogs have been identified among mouse chromosomes (see "The Mighty Mouse," pages 24–25).**


p13.3
p13.2
p12.3
p12.1
p11.2
Breakpoints
STSs
Low-resolution coverage
Mega YACs
High-resolution coverage
Flow-sorted YACs
Contigs
Cosmids
Disease regions
Cloned genes
Genetic markers
Markers homologous to mouse
11
17
Mouse chromosome 16
Mouse chromosome 7

is an especially important step forward, since it is a "sequence-ready" map. It is based on bacterial clones that are ideal substrates for DNA sequencing, and further, these clones have been restriction mapped to allow identification of a minimum set of overlapping clones for a large-scale sequencing effort.

These maps are mere stepping stones to the string of three billion characters – A's, T's, C's, and G's – that defines our species.

The high- and low-resolution maps have been tied together by sequence-tagged sites (STSs), short but unique stretches of DNA sequence. They have also been integrated into the breakpoint map, and with genetic maps developed at the Adelaide Children's Hospital and by CEPH. The integrated map also includes a transcription map of 1000 sequenced *exons* (expressed fragments of genes) and more than 600 other markers developed at other laboratories around the world.

GETTING DOWN TO DETAILS: SEQUENCING THE GENOME

Ultimately, though, these physical maps and the clones they point to are mere stepping stones to the most visible goal of the genome project, the string of three billion characters—A's, T's, C's, and G's—representing the sequence of base pairs that defines our species. Included, of course, would be the sequence for every gene, as well as the sequences for stretches of DNA whose functions we don't yet know (but which may be involved in such little-understood processes as orchestrating gene expression in different parts of our bodies, at different times of our lives). Should anyone undertake to print it all out, the result would fill several hundred volumes the size of a big-city phone book.

Only the barest start has been made in taking this dramatic step in the Human Genome Project. Several hundred million base pairs have been sequenced and archived in databases, but the great majority of these are from short "sequence tags" on cloned fragments. Only about 30 million base pairs of human DNA (roughly one percent of the total) have been sequenced in longer stretches, the longest being about 685,000 base pairs long. Even more daunting is the realization that we will eventually need to sequence many parts of the genome many times, thus to reveal differences that indicate various forms of the same gene.

Hence, as with so many human enterprises, the challenge of sequencing the genome is largely one of doing the job cheaper and faster. At the beginning of the project, the cost of sequencing a single base pair was between $2 and $10, and one researcher could produce between 20,000 and 50,000 base pairs of continuous, accurate sequence in a year. Sequencing the genome by the year 2005 would therefore likely cost $10–20 billion and require a dedicated cadre of at least 5000 workers. Clearly, a major effort in technology development was called for—an effort that would drive the cost well below $1 per base pair and that would allow automation of the sequencing process. From the beginning, therefore, the DOE has emphasized programs to pave the way for expeditious and economical sequencing efforts—programs to develop new technologies, including new cloning vectors, and to establish suitable resources for sequencing, including clone libraries and libraries of expressed sequences.

Efforts to develop new cloning vectors have been especially productive. YACs remain a classic tool for cloning large fragments of human DNA, but they are not perfect. Some regions of the genome, for example, resist cloning in YACs, and others are prone to rearrangement. New vectors such as bacterial artificial chromosomes (BACs), P1 phages, and P1-derived artificial cloning systems (PACs) have thus been devised to address these problems. These new approaches are critical for ensuring that the entire genome can be faithfully represented in clone libraries, without the danger of deletions, rearrangements, or spurious insertions.

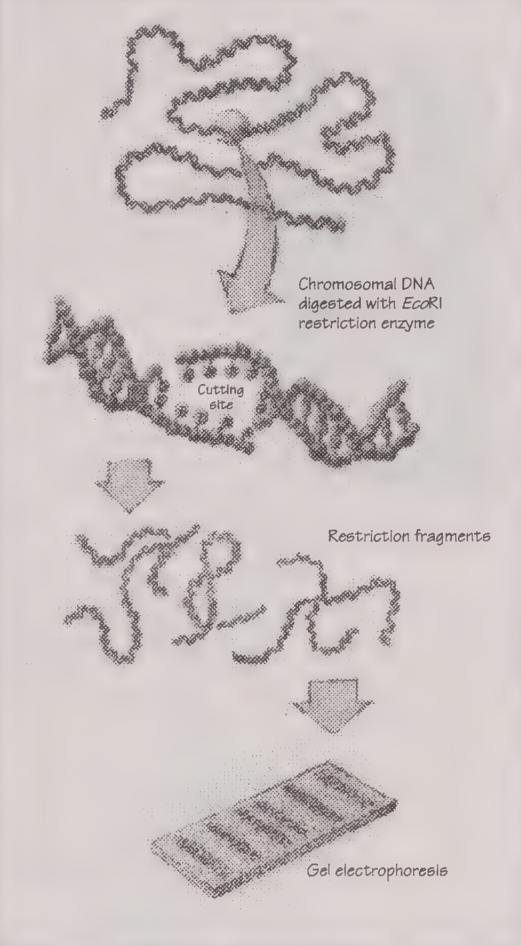
Continues on p. 20

Tools of the Trade

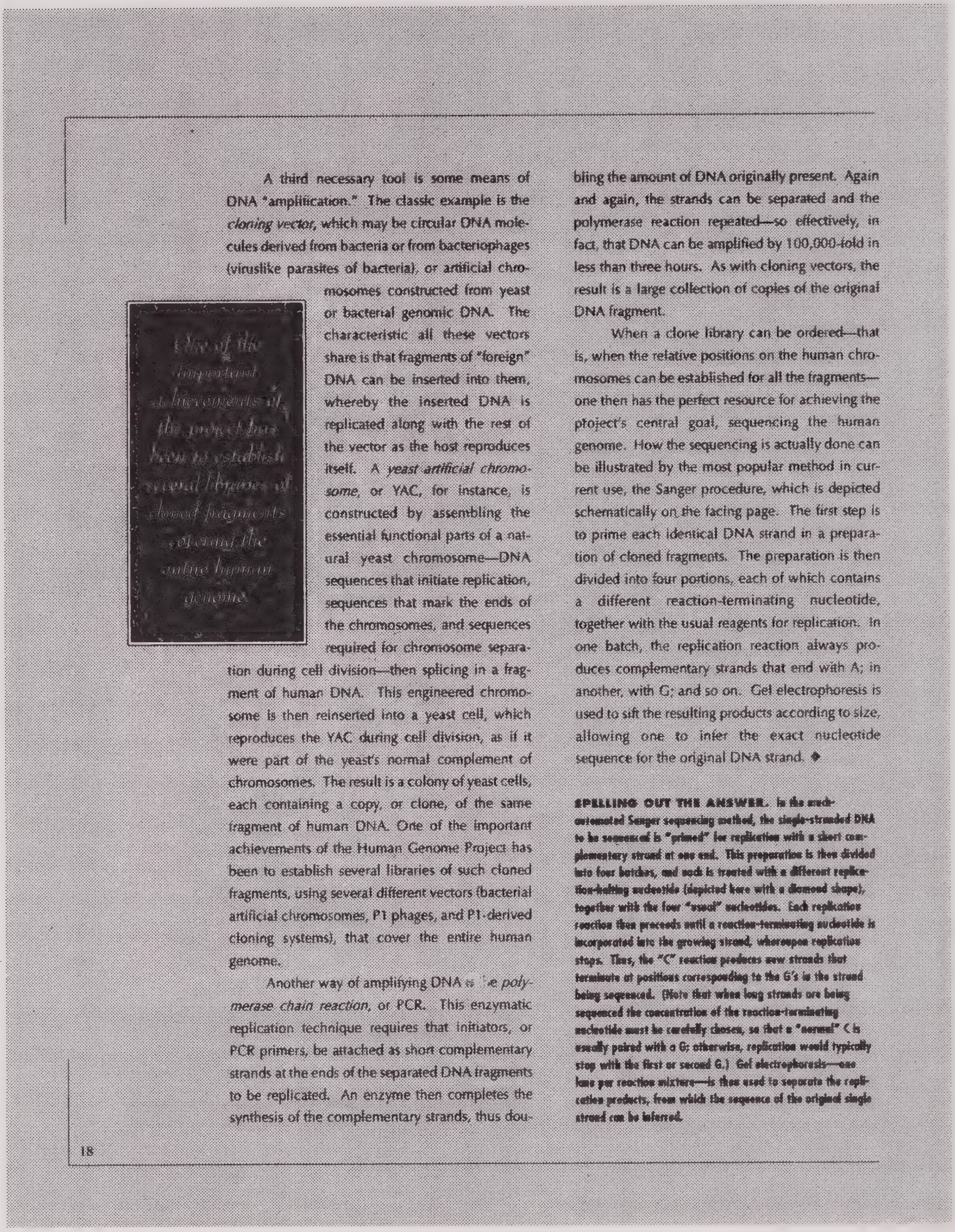
Over the next decade, as molecular biologists tackle the task of sequencing the human genome on a massive scale, any number of innovations can be expected in mapping and sequencing technologies. But several of the central tools of molecular genetics are likely to stay with us—much improved perhaps, but not fundamentally different. One such tool is the class of DNA-cutting proteins known as *restriction enzymes*. These enzymes, the first of which were discovered in the late 1960s, cleave double-stranded DNA molecules at specific recognition sites, usually four or six nucleotides long. For example, a restriction enzyme called *Eco*RI recognizes the single-strand sequence GAATTC and invariably cuts the double helix as shown in the illustration on the right.

When digested with a particular restriction enzyme, then, identical segments of human DNA yield identical sets of restriction fragments. On the other hand, DNA from the same genomic region of two different people, with their subtly different genomic sequences, can yield dissimilar sets of fragments, which then produce different patterns when sorted according to size.

This leads directly to discussion of a second essential tool of modern molecular genetics, *gel electrophoresis*, for it is by electrophoresis that DNA fragments of different sizes are most often separated. In classical gel electrophoresis, electrically charged macromolecules are caused to migrate through a polymeric gel under the influence of an imposed static electric field. In time the molecules sort themselves by size, since the smaller ones move more rapidly through the gel than do larger ones. In 1984 a further advance was made with the invention of pulsed-field gel electrophoresis, in which the strength and direction of the applied field is varied rapidly, thus allowing DNA strands of more than 50,000 base pairs to be separated.



DIGESTING DNA. Isolated from various bacteria, restriction enzymes serve as microscopic scalpels that cut DNA molecules at specific sites. The enzyme *Eco*RI, for example, cuts double-stranded DNA only where it finds the sequence GAATTC. The resulting fragments can then be separated by gel electrophoresis. The electrophoresis pattern itself can be of interest, since variations in the pattern from a given chromosomal region can sometimes be associated with variations in genetic traits, including susceptibilities to certain diseases. Knowledge of the cutting sites also yields a kind of physical map known as a restriction map.



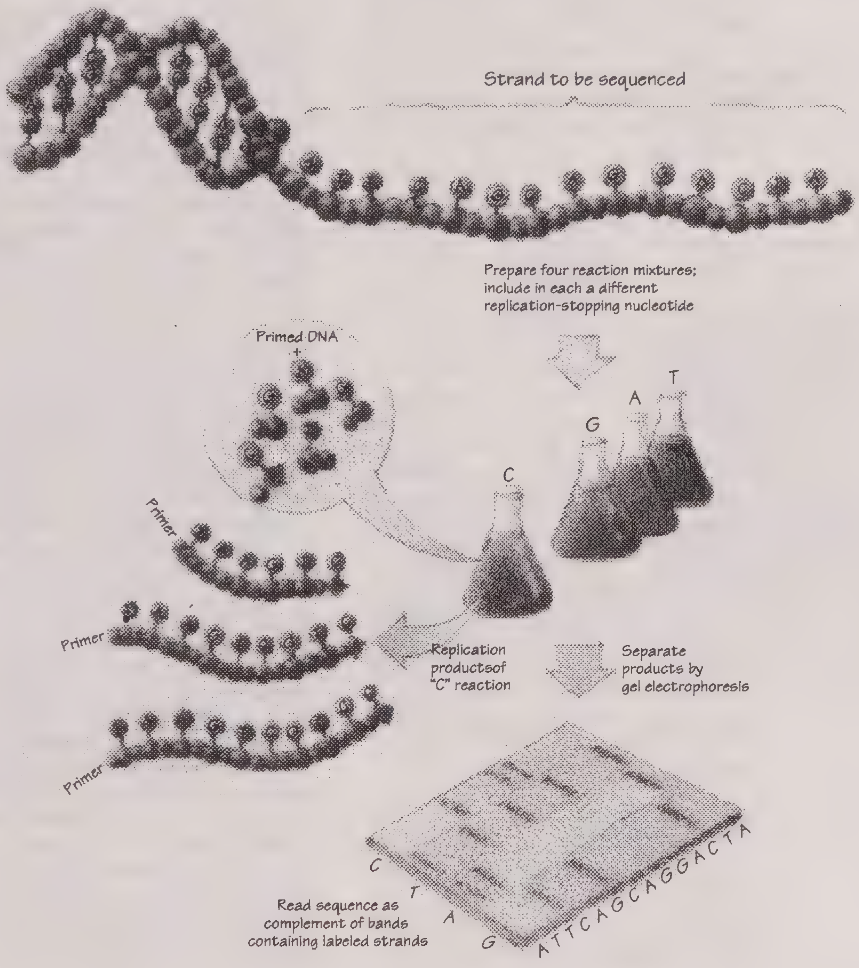
A third necessary tool is some means of DNA "amplification." The classic example is the *cloning vector*, which may be circular DNA molecules derived from bacteria or from bacteriophages (viruslike parasites of bacteria), or artificial chromosomes constructed from yeast or bacterial genomic DNA. The characteristic all these vectors share is that fragments of "foreign" DNA can be inserted into them, whereby the inserted DNA is replicated along with the rest of the vector as the host reproduces itself. A *yeast artificial chromosome*, or YAC, for instance, is constructed by assembling the essential functional parts of a natural yeast chromosome—DNA sequences that initiate replication, sequences that mark the ends of the chromosomes, and sequences required for chromosome separation during cell division—then splicing in a fragment of human DNA. This engineered chromosome is then reinserted into a yeast cell, which reproduces the YAC during cell division, as if it were part of the yeast's normal complement of chromosomes. The result is a colony of yeast cells, each containing a copy, or clone, of the same fragment of human DNA. One of the important achievements of the Human Genome Project has been to establish several libraries of such cloned fragments, using several different vectors (bacterial artificial chromosomes, P1 phages, and P1-derived cloning systems), that cover the entire human genome.

One of the important achievements of the project has been to establish several libraries of cloned fragments covering the entire human genome.

Another way of amplifying DNA is the *polymerase chain reaction*, or PCR. This enzymatic replication technique requires that initiators, or PCR primers, be attached as short complementary strands at the ends of the separated DNA fragments to be replicated. An enzyme then completes the synthesis of the complementary strands, thus doubling the amount of DNA originally present. Again and again, the strands can be separated and the polymerase reaction repeated—so effectively, in fact, that DNA can be amplified by 100,000-fold in less than three hours. As with cloning vectors, the result is a large collection of copies of the original DNA fragment.

When a clone library can be ordered—that is, when the relative positions on the human chromosomes can be established for all the fragments—one then has the perfect resource for achieving the project's central goal, sequencing the human genome. How the sequencing is actually done can be illustrated by the most popular method in current use, the Sanger procedure, which is depicted schematically on the facing page. The first step is to prime each identical DNA strand in a preparation of cloned fragments. The preparation is then divided into four portions, each of which contains a different reaction-terminating nucleotide, together with the usual reagents for replication. In one batch, the replication reaction always produces complementary strands that end with A; in another, with G; and so on. Gel electrophoresis is used to sift the resulting products according to size, allowing one to infer the exact nucleotide sequence for the original DNA strand. ◆

SPELLING OUT THE ANSWER. **In the much-automated Sanger sequencing method, the single-stranded DNA to be sequenced is "primed" for replication with a short complementary strand at one end. This preparation is then divided into four batches, and each is treated with a different replication-halting nucleotide (depicted here with a diamond shape), together with the four "usual" nucleotides. Each replication reaction then proceeds until a reaction-terminating nucleotide is incorporated into the growing strand, whereupon replication stops. Thus, the "C" reaction produces new strands that terminate at positions corresponding to the G's in the strand being sequenced. (Note that when long strands are being sequenced the concentration of the reaction-terminating nucleotide must be carefully chosen, so that a "normal" C is usually paired with a G; otherwise, replication would typically stop with the first or second G.) Gel electrophoresis—one lane per reaction mixture—is then used to separate the replication products, from which the sequence of the original single strand can be inferred.**


Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA
+
G
A
T
C
Primer
Primer
Primer
Replication productsof "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read sequence as complement of bands containing labeled strands

Marked progress is also evident in the development of sequencing technologies, though all of those in widespread current use are still based on methods developed in 1977 by Allan Maxam and Walter Gilbert and by Frederick Sanger and his coworkers (see "Tools of the Trade," pages 17–19). Both of these methods rely on gel-based electrophoresis systems to separate DNA fragments, and recent advances in commercial systems include increasing the number of gel lanes, decreasing run times, and enhancing the accuracy of base identification. As a result of such improvements, a standard sequencing machine can now turn out raw, unverified sequences of 50,000 to 75,000 bases per day.

Equally important to the sequencing goals of the genome project is a rational system for organizing and distributing the material to be sequenced. The DOE's commitment to such resources dates back to 1984, when it organized the National Laboratory Gene Library Project. Based on cell- and chromosome-sorting technologies developed at Livermore and Los Alamos, libraries of clones were established for each of the human chromosomes, and the individual clones are widely available for mapping and for isolating genes. These clones were invaluable in such notable "gene hunts" as the successful searches for the cystic fibrosis and Huntington disease genes. More recently, as more efficient vectors have become available, complete human DNA libraries have been established using BACs, PACs, and YACs.

> *Advances have brought much nearer the day when "production sequencing" can begin.*

Another critical resource is being assembled in an effort known as I.M.A.G.E. (Integrated Molecular Analysis of Genomes and their Expression), cofounded by the Livermore Human Genome Center. The aim is a master set of mapped and sequenced human cDNA, representing the expressed parts of the human genome. By early 1996, I.M.A.G.E. had distributed over 250,000 partial and complete cDNA clones, most of them with one or both ends sequenced to provide unique identifiers. These identifiers, *expressed sequence tags* (ESTs), are usually 300–500 base pairs each. Twenty-five hundred genes have also been newly mapped as part of this coordinated effort.

SHOTGUNS AND TRANSPOSONS

Such advances as these, in both technology development and the assembly of resource libraries, have brought much nearer the day when "production sequencing" can begin. A great deal of variety remains, however, in the approaches available to sequencing the human genome, and it is not yet clear which will prove the most efficient and most cost-effective way to read long stretches of DNA over the next decade. One of the available choices, for example, is between "shotgun" and "directed" strategies. Another is the degree of redundancy—that is, how many times must a given strand be sequenced to ensure acceptable confidence in the result?

Shotgun sequencing derives its name from the randomly generated DNA fragments that are the objects of scrutiny. Many copies of a single large clone are broken into pieces of perhaps 1500 base pairs, either by restriction enzymes or by physical shearing. Each fragment is then separately cloned, and a convenient portion of it sequenced. A computational assembly process then compares the terminal sequences of the many fragments and, by finding overlaps that indicate neighboring fragments, constructs an ordered library for the parent clone. The members of this ordered library can then be sequenced from end to end to yield a complete sequence for the parent. The statistics involved in taking this approach require that many copies of the original clone be randomly fragmented, if no gaps are to be tolerated in the final sequence. A benefit is that the final sequence is highly reliable; the main disadvantage is that the same sequence must be done many times (in the many overlapping fragments). Nevertheless, shotgun sequencing has been the primary means for generating most of the genomic sequence data in public DNA databases. This includes the longest contiguous fragment of sequenced human DNA, from the human T-cell receptor beta region, of about 685,000 base pairs—a product of DOE-supported work at the University of Washington.

The shotgun strategy is also being used at the Genome Therapeutics Corporation and The Institute for Genomic Research (TIGR), as part of the DOE-supported Microbial Genome Initiative. Genome Therapeutics has sequenced 1.8 million base pairs of *Methanobacterium thermoautotrophicum*, a bacterium important in energy production and bioremediation, and TIGR has successfully sequenced the complete genomes of three free-living bacteria, *Haemophilus influenzae* (1,830,137 base pairs; an effort supported mostly by private funds), *Mycoplasma genitalium* (580,070 base pairs), and *Methanococcus jannaschii* (1,739,933 base pairs).

The alternative to shotgun sequencing is a directed approach, in which one seeks to sequence the target clone from end to end with a minimum of duplication. The essence of this approach is embodied in a technique known as *primer walking*. Starting at one end of a single large fragment, one replicates a stretch of DNA—say, 400 base pairs long—that can be sequenced in one run. With the sequence for this first segment in hand, the next stretch of DNA, just overlapping the first, is then tackled in the same way. In principle, one can thus "walk" the entire length of the original clone. Unfortunately, this conceptually simple approach has been historically beset with disadvantages, mainly the expense and inconvenience of custom-synthesizing a primer as the necessary starting point for each sequencing step. The widely automated Sanger sequencing method involves a DNA replication step that must be "primed" by a DNA fragment that is complementary to 15 to 20 base pairs of the strand to be sequenced (see "Tools of the Trade," pages 17–19). Until recently, making these primers was an expensive and time-consuming business, but recent innovations have made primer walking, and similar directed strategies, more and more economically feasible.

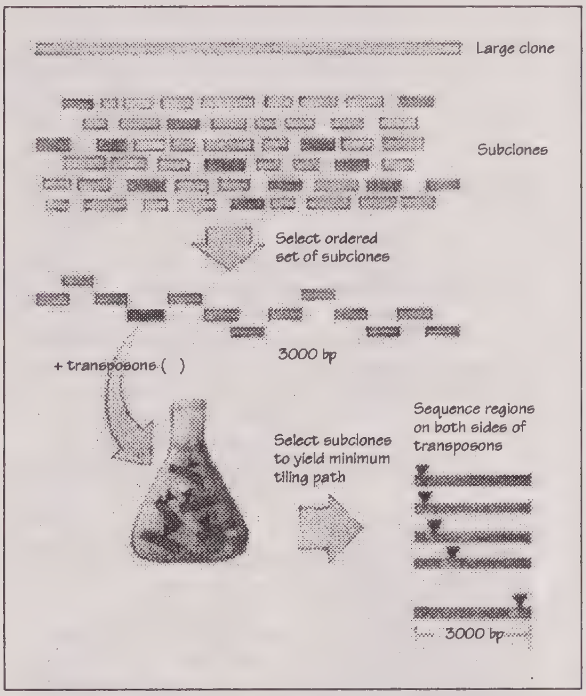

Large clone
Subclones
Select ordered set of subclones
3000 bp
+ transposons ( )
Select subclones to yield minimum tiling path
Sequence regions on both sides of transposons
3000 bp

FIGURE 7. TAKING A DIRECTED APPROACH. One directed sequencing strategy exploits a naturally occurring genetic element known as a transposon. The starting point is an ordered set of subclones, each about 3000 base pairs long, derived from a much larger clone (say, a YAC). For each subclone, a preparation is then made in which transposons insert themselves randomly into the subclone—on average, one transposon in each 3000-base-pair strand. The positions of the transposons are mapped, and a set of strands is selected such that the insertion points are about 300 base pairs apart. Sequencing then proceeds in both directions from the transposon insertion points, using the known transposon sequence as a primer. The full set of overlapping regions yields the sequence for the entire subclone, and the sequences of the full set of subclones yield the sequence for the larger original clone.

One way to deal with the primer bottleneck, for example, is to use sets of very short fragments to prime the next sequencing step. As an illustration, the four nucleotides (A, T, C, and G) can be ordered in more than 68 billion ways to create an 18-base primer, an imposing set of possibilities. But it is eminently practical to create a library of the 4096 possible 6-base primers. Three of these "6-mers" can be matched to the end of the fragment to be sequenced, thus serving as an 18-base primer. This modular primer technology, developed at the Brookhaven National Laboratory, is currently being applied to *Borrelia burgdorferi*, the organism that causes Lyme disease; a 34,000-base-pair fragment has already been sequenced.

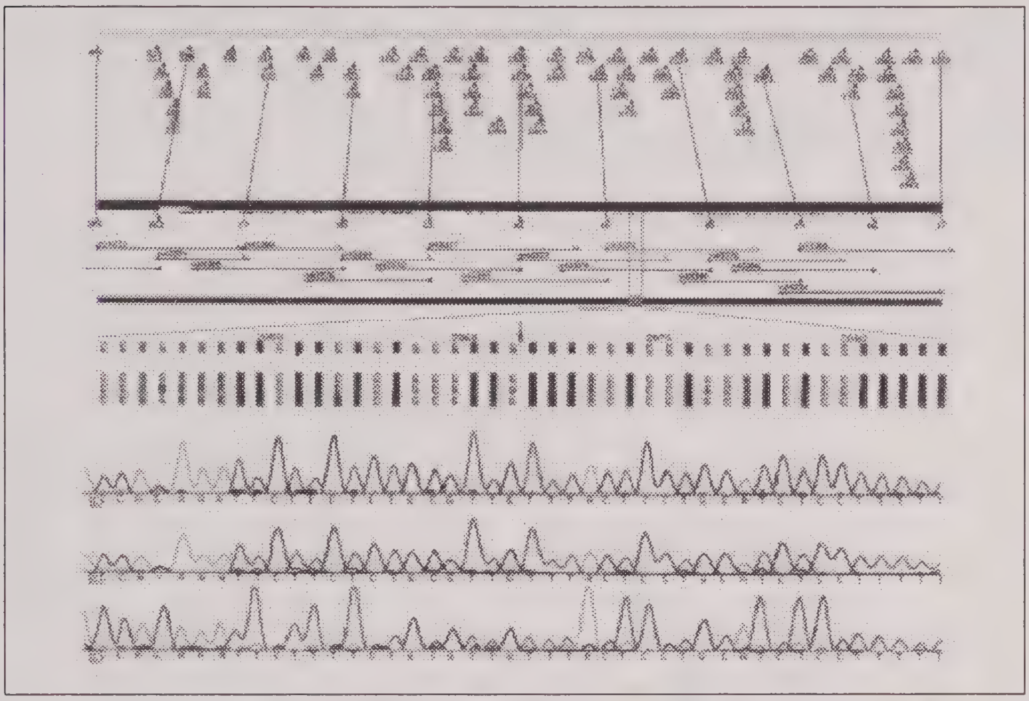
Another directed approach uses a naturally occurring genetic element called a *transposon*, which insinuates itself more or less randomly in longer DNA strands. This predilection for random insertion and the fact that the transposon's DNA sequence is well known are the keys to the sequencing strategy depicted schematically in Figure 7. The largest clones are broken into smaller subclones (each of about 3000 base pairs), which then become the targets of the transposons. Multiple copies of each subclone are exposed to the transposons, and reaction conditions are controlled to yield, on average, a single insertion in each 3000-base-pair strand. The individual strands are then analyzed to yield, for each, the approximate position of the inserted transposon. By mapping these positions, a "minimum tiling path" can be determined for each subclone—that is, a set of strands can be identified whose transposon insertions are roughly 300 base pairs apart. In this set of strands, the region around each transposon is then sequenced, using the inserted transposons as starting points. The known transposon sequence allows a single primer to be used for sequencing the full set of overlapping regions.

The completed physical maps of chromosomes 16 and 19 are especially ripe for large-scale sequencing.

At the Lawrence Berkeley National Laboratory, this technique has been used to sequence over 1.5 million base pairs of DNA on human chromosomes 5 and 20, as well as over three million base pairs from the fruit fly *Drosophila melanogaster*. On chromosome 5, interest focuses on a region of three million base pairs that is rich in growth factor and receptor genes; whereas, on chromosome 20, Berkeley researchers are interested in a region of about two million base pairs that is implicated in 15 to 20 percent of all primary breast carcinomas. As an example of the kind of output these efforts produce, Figure 8 shows a stretch of sequence data from chromosome 5.

Researchers supported by the DOE at the University of Utah are also pursuing the use of directed sequencing. In addition, they have developed a methodology for "multiplex" DNA sequencing, which offers a way of increasing throughput with either shotgun or directed approaches. By attaching a unique identifying sequence to each sequencing sample in a mixture of, say, 50 such samples, the entire mixture can be analyzed in a single electrophoresis lane. The 50 samples can be resolved sequentially by probing, first, for bands containing the first identifier, then for bands containing the second, and so forth. In a similar way, multiplexing can also be used for mapping. The Utah group is now able to map almost 5000 transposons in a single experiment, and they are using multiplexing in concert with a directed sequencing strategy to sequence the 1.8 million base pairs of the thermophilic microbe *Pyrococcus furiosus* and two important regions of human chromosome 17.

The completed physical maps of chromosomes 16 and 19, with their extensive coverage in many different kinds of cloning vectors, are especially ripe for large-scale sequencing. Los Alamos scientists have therefore begun sequencing chromosome 16, focusing special effort on locating the estimated 3000 expressed genes on that chromosome and using those sites as starting points for directed genomic sequencing. A region of 60,000 base pairs has already been sequenced around the adult polycystic kidney gene, and good starts have been made in mapping other genes. Interestingly, even random sequencing has led to the identification of gene DNA in over 15 percent of the samples, confirming the apparent high density of genes on this chromosome. Between chromosome 16 and the short arm of chromosome 5, another Los Alamos target, the genome center there



has produced almost two million base pairs of human DNA sequence.

A parallel effort is under way at Livermore on chromosome 19 and other targeted genomic regions. Using a shotgun approach, researchers there have completed over 1.3 million bases of genomic sequence. Initially, they are attacking two major regions of chromosome 19: one of about two million base pairs, containing several genes involved in DNA repair and replication, and another of approximately one million base pairs, containing a kidney disease gene. The Livermore scientists are making use of the I.M.A.G.E. cDNA resource to sequence the cDNA from these regions, along with the associated segments of the genome. In addition, Livermore scientists have targeted DNA repair gene regions throughout the genome and, in many cases, have done comparative sequencing of these genes in other

Continues on p. 26

FIGURE 8. SEQUENCE DATA: THE FINAL PRODUCT. The ultimate description of the genome, though only a prelude to full understanding, is the base-pair sequence. This computer display shows results from the use of transposons at Berkeley. The array of triangles represents the transposons inserted into a 3000-base-pair subclone; the 11 selected by the computer to build a minimum tiling path are shown below the heaviest black line. The subclone segments sequenced by using these 11 starting points are depicted by the horizontal lines; the arrowheads indicate the sequencing directions. The expanded region between bases 2042 and 2085 is covered by three sequencing reactions, which produced the three traces at the bottom of the figure. Above the traces, the results are summarized, together with a consensus sequence (just below the numbers).

The Mighty Mouse

The human genome is not so very different from that of chimpanzees or mice, and it even shares many common elements with the genome of the lowly fruit fly. Obviously, the differences are critical, but so are the similarities. In particular, genetic experiments on other organisms can illuminate much that we could not otherwise learn about *homologous* human genes—that is, genes that are basically the same in the two species.

In some cases, the connection between a newly identified human gene and a known health disorder can be quickly established. More often, however, clear links between cloned genes and human hereditary diseases or disease susceptibilities are extremely elusive. Diseases that are modified by other genetic predispositions, for example, or by environment, diet, and lifestyle can be exceedingly difficult to trace in human families. The same holds for very rare diseases and for genetic factors contributing to birth defects and other developmental disorders. By contrast, disorders such as these can sometimes be followed relatively easily in animal systems, where uniform genetic backgrounds and controlled breeding schemes can be used to avoid the variability that often confounds human population studies. As a consequence, researchers looking for clues to the causes of many complex health problems are focusing more and more attention on model animal systems.

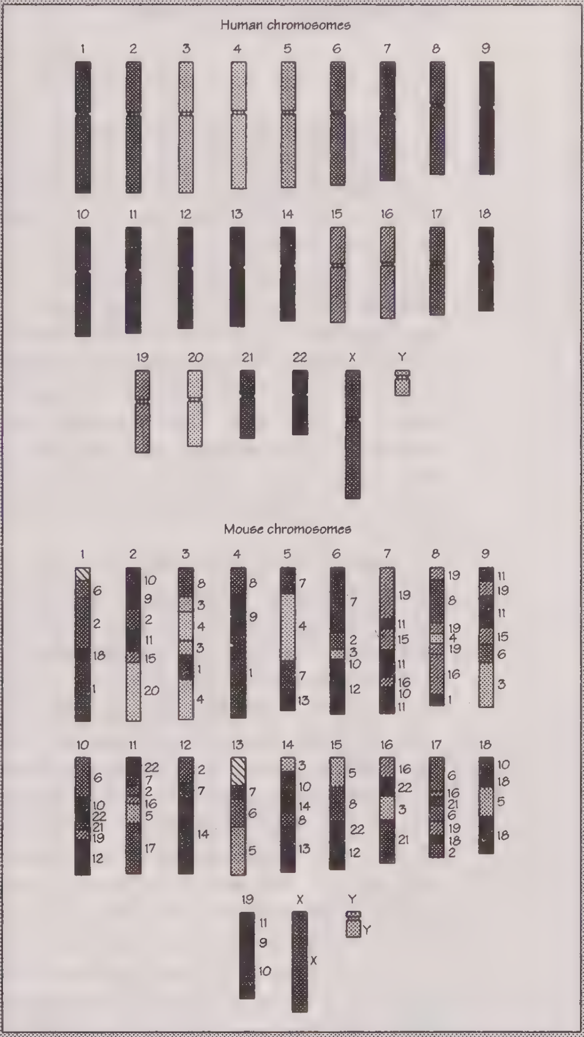
Among such systems, which range in complexity from yeast and bacteria to mammals, the most prominent is the mouse. Because of its small size, high fertility rate, and experimental manipulability, the mouse offers great promise in studying the genetic causes and pathological progress of ailments, as well as understanding the genetic role in disease susceptibility. In pursuing such studies, the DOE is exploiting several resources, among them the experimental mouse genetics facility at the Oak Ridge National Laboratory. Initially established for genetic risk assessment and toxicology studies, the Oak Ridge facility is one of the world's largest. Mutant strains there express a variety of inherited developmental and health disorders, ranging from dwarfism and limb deformities to sickle cell anemia, atherosclerosis, and unusual susceptibilities to cancer.

Most of these existing mutant strains have arisen from random alterations of genes, caused by the same processes that occur naturally in all living populations. However, other, more directed means of gene alteration are also available. So-called *transgenic* methods, which have been developed and refined over the past 15 years, allow DNA sequences engineered in the laboratory to be introduced directly into the genomes of mouse embryos. The embryos are subsequently transferred to a foster mother, where they develop into mice carrying specifically designed alterations in a particular gene. The differences in form, basic health, fertility, and longevity produced by these "designer mutations" then allow researchers to study the effects of genetic defects that can mimic those found in human patients. The payoff can be clues that aid in the design of drugs and other treatments for the human diseases.

The Human Genome Center at Berkeley is using mice for similar purposes. In vivo libraries of overlapping human genome fragments (each 100,000 to 1,000,000 base pairs long) are being propagated in transgenic mice. The region of chromosome 21 responsible for Down syndrome, for example, is now almost fully represented in a panel of transgenic mice. Such libraries have several uses. For example, the precise biochemical means by which identified genes produce their effects can be studied in detail, and new genes can be recognized by analyzing the effects of particular genome fragments on the transgenic animals. In such ways, the promise of the massive effort to map and sequence the human genome can be translated into the kind of biological knowledge coveted by pharmaceutical designers and medical researchers.

Adding to the potential value of mutant mice as models for human genetic disease is growing evidence of similarities between mouse and human genes. Indeed, practically every human gene appears to have a counterpart in the mouse genome. Furthermore, the related mouse and human genes often share very similar DNA sequences and the same basic biological function. If we imagine that the 23 pairs of human chromosomes were shattered into smaller blocks—to yield a total of, say, 150 pieces, ranging in size from very small bits containing just a few genes to whole chromosome arms—those pieces could be reassembled to produce a serviceable model of the mouse genome. This mouse genome jigsaw puzzle is shown to the right. Thanks to this mouse-human genomic homology, a newly located gene on a human chromosome can often lead to a confident prediction of where a closely related gene will be found in the mouse—and vice versa.

Thus, a crippling heritable muscle disorder in mice maps to a location on the mouse X chromosome that is closely analogous to the map location for the X-linked human Duchenne muscular dystrophy gene (DMD). Indeed, we now know that these two similar diseases are caused by the mouse and human versions of the same gene. Although mutations in the mouse *mdx* gene produce a muscle disease that is less severe than the heartbreaking, fatal disease resulting from the DMD mutation in humans, the two genes produce proteins that function in very similar ways and that are clearly required for normal muscle development and function in the corresponding species. Likewise, the discovery of a mouse gene associated with pigmentation, reproductive, and blood cell defects was the crucial key to uncovering the basis for a human disease known as the piebald trait. Owing to such close human-mouse relationships as these, together with the benefits of transgenic technologies, the mouse offers enormous potential in identifying new human genes, deciphering their complex functions, and even treating genetic diseases. ◆



OF MICE AND MEN. The genetic similarity (or homology) of superficially dissimilar species is amply demonstrated here. The full complement of human chromosomes can be cut, schematically at least, into about 150 pieces (only about 100 are large enough to appear in this illustration), then reassembled into a reasonable approximation of the mouse genome. The colors of the mouse chromosomes and the numbers alongside indicate the human chromosomes containing homologous segments. This piecewise similarity between the mouse and human genomes means that insights into mouse genetics are likely to illuminate human genetics as well.

species, especially the mouse. Such comparative sequencing has identified conserved sequence elements that might act as regulatory regions for these genes and has also assisted in the identification of gene function (see "The Mighty Mouse," pages 24–25).

HOW GOOD IS GOOD ENOUGH?

The goal of most sequencing to date has been to guarantee an error rate below 1 in 10,000, sometimes even 1 in 100,000. However, the difference between one human being and another is more like one base pair in five hundred, so most researchers now agree that one error in a thousand is a more reasonable standard. To assure a higher level of confidence, and perhaps to uncover important individual differences, the most biologically or medically important regions would still be sequenced more exhaustively, but using this lowered standard would greatly reduce the cost of acquiring sequence data for the bulk of human DNA.

With this philosophy in mind, Los Alamos scientists have begun a project to determine the cost and throughput of a low-redundancy sequencing strategy known as *sample sequencing* (SASE, or "sassy"). Clones are selected from the high-resolution Los Alamos cosmid map, then physically broken into 3000-base-pair subclones—much as in other sequencing approaches. In contrast to, say, shotgun sequencing, though, only a small random set of the subclones is then selected for sequencing. Sequence fragments already known—end sequences, sequence-tagged sites, and so forth—are used as the starting points. The result is sequence coverage for about 70 percent of the original cosmid clone, enough to allow identification of genes and ESTs, thus pinpointing the most critical targets for later, more thorough sequencing efforts. Further, the SASE-derived sequences provide enough information for researchers elsewhere to pursue just such comprehensive efforts, using whole genomic DNA. In addition, the cost of SASE sequencing is only one-tenth the cost of obtaining a complete sequence, and a genomic region can be "sampled" ten times as fast.

As the first major target of SASE analysis, Los Alamos scientists chose a cosmid contig of four million base pairs at the end (the *telomere*) of the short arm of chromosome 16. By early 1996, over 1.4 million base pairs had been sequenced, and a gene, EST, or suspected coding region had been located on every cosmid sampled.

In addition, Los Alamos is building on the SASE effort by using SASE sequence data as the basis for an efficient primer walking strategy for detailed genomic sequencing. The first application of this strategy, to a telomeric region on the long arm of chromosome 7, proved to be as efficient as typical shotgun sequencing, but it required only two- to threefold redundancy to produce a complete sequence, in contrast to the seven- to tenfold redundancy required in shotgun approaches. The resulting 230,000-base-pair sequence is the second-longest stretch of contiguous human DNA sequence ever produced.

◆

In a sense, though, even a complete genome sequence—the ultimate physical map—is only a start in understanding the human genome. The deepest mystery is how the potential of 100,000 genes is regulated and controlled, how blood cells and brain cells are able to perform their very different functions with the same genetic program, and how these and countless other cell types arise in the first place from an single undifferentiated egg cell. A first step toward solving these subtle mysteries, though, is a more complete physical picture of the master molecules that lie at the heart of it all.

Beyond Biology

INSTRUMENTATION AND INFORMATICS

FROM THE START, it has been clear that the Human Genome Project would require advanced instrumentation and automation if its mapping and sequencing goals were to be met. And here, especially, the DOE's engineering infrastructure and tradition of instrumentation development have been crucial contributors to the international effort. Significant DOE resources have been committed to innovations in instrumentation, ranging from straightforward applications of automation to improve the speed and efficiency of conventional laboratory protocols (see, for example, Figure 9a) to the development of technologies on the cutting edge—technologies that might potentially increase mapping and sequencing efficiencies by orders of magnitude.

On the first of these fronts, genome researchers are seeing significant improvements in the rate, efficiency, and economy of large-scale mapping and sequencing efforts as a result of improved laboratory automation tools. In many cases, commercial robots have simply been mechanically reconfigured and reprogrammed to perform repetitive tasks, including the replication of large clone libraries, the pooling of libraries as a prelude to various assays, and the arraying of clone libraries for hybridization studies. In other cases, custom-designed instruments have proved more efficient. A notable illustration is the world's fastest cell and chromosome sorter, developed at Livermore and now being commercialized, which is used to sort human chromosomes for chromosome-specific libraries. Other examples include a high-speed, robotics-compatible thermal cycler developed at Berkeley, which greatly accelerates PCR amplifications, and instruments developed at Utah for automated hybridization in multiplex sequencing schemes.

SMALLER IS BETTER—AND OTHER DEVELOPMENTS

Beyond "mere" automation are efforts aimed at more fundamental enhancements of established techniques. In particular, a number of DOE-supported efforts aim at improved versions of the automated gel-based Sanger sequencing technique. For example, in place of the conventional slab gels, ultrathin gels, less than 0.1 millimeter thick, can be used to obtain 400 bases of sequence from each lane in a hour's run, a fivefold improvement in throughput over conventional systems. Even faster speedups are seen when arrays of 0.1-millimeter capillaries are used as the separation medium. Both of these approaches exploit higher electric field strengths to increase DNA mobility and to reduce analysis times. And Livermore scientists are looking beyond even capillaries, to sequencing arrays of rigid glass microchannels, supplemented by automated gel and sample loading.

The project will require advanced instrumentation and automation if its goals are to be met.

The capillary approach is especially ripe for further development. Challenges include providing uniform excitation over

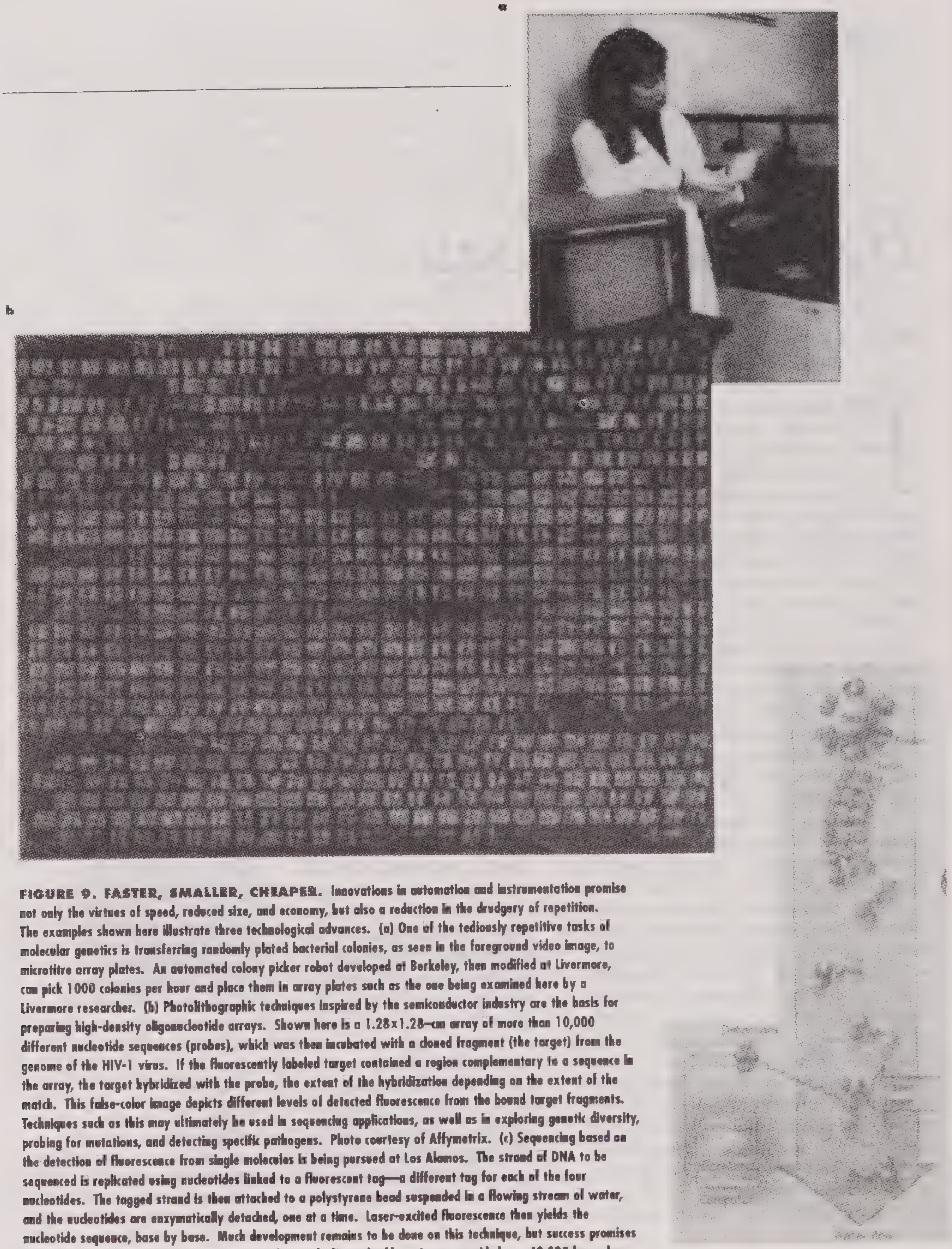


FIGURE 9. FASTER, SMALLER, CHEAPER. Innovations in automation and instrumentation promise not only the virtues of speed, reduced size, and economy, but also a reduction in the drudgery of repetition. The examples shown here illustrate three technological advances. (a) One of the tediously repetitive tasks of molecular genetics is transferring randomly plated bacterial colonies, as seen in the foreground video image, to microtitre array plates. An automated colony picker robot developed at Berkeley, then modified at Livermore, can pick 1000 colonies per hour and place them in array plates such as the one being examined here by a Livermore researcher. (b) Photolithographic techniques inspired by the semiconductor industry are the basis for preparing high-density oligonucleotide arrays. Shown here is a 1.28×1.28-cm array of more than 10,000 different nucleotide sequences (probes), which was then incubated with a cloned fragment (the target) from the genome of the HIV-1 virus. If the fluorescently labeled target contained a region complementary to a sequence in the array, the target hybridized with the probe, the extent of the hybridization depending on the extent of the match. This false-color image depicts different levels of detected fluorescence from the bound target fragments. Techniques such as this may ultimately be used in sequencing applications, as well as in exploring genetic diversity, probing for mutations, and detecting specific pathogens. Photo courtesy of Affymetrix. (c) Sequencing based on the detection of fluorescence from single molecules is being pursued at Los Alamos. The strand of DNA to be sequenced is replicated using nucleotides linked to a fluorescent tag—a different tag for each of the four nucleotides. The tagged strand is then attached to a polystyrene bead suspended in a flowing stream of water, and the nucleotides are enzymatically detached, one at a time. Laser-excited fluorescence then yields the nucleotide sequence, base by base. Much development remains to be done on this technique, but success promises a cheaper, faster approach to sequencing, one that might be applicable to intact cosmid clones 40,000 bases long.

arrays of 50 to 100 capillaries and then efficiently detecting the fluorescence emitted by labeled samples. Technologies under investigation include fiber-optic arrays, scanning confocal microscopy, and cooled CCD cameras. Some of this effort has already been transferred to the private sector, and tenfold improvements in speed, economy, and efficiency are projected in future commercial instruments.

The move toward miniaturization is afoot elsewhere as well. Building on experiences in the electronics industry, several DOE-supported groups are exploring ways to adapt high-resolution photolithographic methods to the manipulation of minuscule quantities of biological reagents, followed by assays performed on the same "chip." Current thrusts of this "nanotechnology" approach include the design of microscopic electrophoresis systems and ultrasmall-volume, high-speed thermal cycling systems for PCR. A miniaturized, computer-controlled PCR device under development at Livermore operates on 9-volt batteries and might ultimately lead to arrays of thousands of individually controlled micro-PCR chambers.

Another miniaturization effort aims at the fabrication of high-density combinatorial arrays of custom *oligomers* (short chains of nucleotides), which would make feasible large-scale hybridization assays, including sequencing by hybridization. This innovative technique uses short oligomers that pair up with corresponding sequences of DNA. The oligomers are placed on an array by a process similar to that of making silicon chips for electronics. Successful matches between oligomers and genomic DNA are then detected by fluorescence, and the application of sophisticated statistical analyses reassembles the target sequence. This same technology has already been used for genetic screening and cDNA fingerprinting. Figure 9b illustrates a DOE-supported application of high-density oligonucleotide arrays to the detection of mutations in the HIV-1 genome. Similar approaches can be envisioned to understand differences in patterns of gene expression: Which genes are active (which are producing mRNA) in which cells? Which are active at different times during an organism's development? Which are active, or inactive, in disease?

Sequencing by hybridization is only one of several forward-looking ideas for revolutionizing sequencing technology. In spite of continuing improvements to sequencers based on the classic methods, it is nonetheless desirable to explore altogether new approaches, with an eye to simplifying sample preparation, reducing measurement times, increasing the length of the strands that can be analyzed in a single run, and facilitating interpretation of the results. Over the course of the past few years, several alternative approaches to direct sequencing have been explored, including atomic-resolution molecular scanning, single-molecule detection of individual bases, and mass spectrometry of DNA fragments.

In spite of improvements to sequencers based on the classic methods, it is nonetheless desirable to explore altogether new approaches.

All of these alternatives look promising in the long term, but mass spectrometry has perhaps demonstrated the greatest near-term potential. Mass spectrometry measures the masses of ionized DNA fragments by recording their time-of-flight in vacuum. It would therefore replace traditional gel electrophoresis as the last step in a conventional sequencing scheme. Routine application of this technique still lies in the future, but fragments of up to 500 bases have been analyzed, and practical systems based on high-resolution mass separations of DNA fragments of fewer than 100 bases are currently being developed at several universities and national laboratories.

Another innovative sequencing method is under investigation at Los Alamos. As depicted in Figure 9c, each of the four bases (A, T, C, G) in a single strand of DNA receives a different fluorescent label, then the bases are enzymatically detached, one at a time. The characteristic fluorescence is detected by a laser system, thereby yielding the sequence, base by base. This approach is beset by major technical challenges, and direct

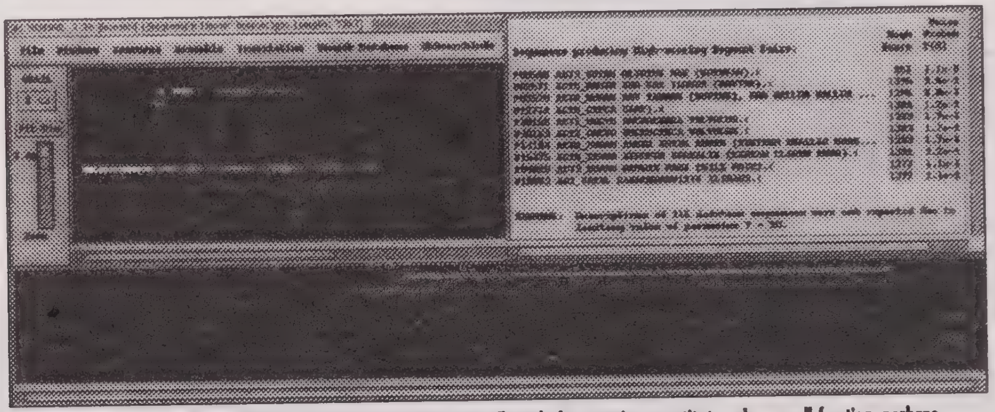


FIGURE 10. GENE HUNTS. Genes, the regions that actually code for proteins, constitute only a small fraction, perhaps 10%, of the human genome. Thus, even with sequence in hand, finding the genes is yet another daunting step away. One tool developed to help in the hunt is GRAIL, a computer program developed at Oak Ridge that uses heuristics based on existing data, together with artificial neural networks, to identify likely genes. Coding and noncoding regions of the genome differ in many subtle respects—for example, the frequency with which certain short sequences appear. Further, particular landmarks are known to characterize the boundaries of many genes. In the example shown here, GRAIL has searched for likely genes in both strands of a 3583-base-pair sequence. The results are shown at the upper left. The upper white trace indicates five possible exons (coding regions within a single gene) in one strand, whereas the lower white trace suggests two possible exons in the other strand. However, the lower trace scores worse on other tests, leading to a candidate set of exons shown by the five green rectangles. By refining this set further, GRAIL then produces the final gene model shown in light blue. The lower part of the figure zeros in on the end of the candidate exon outlined in yellow, thus providing a detailed look at one of the differences between the preliminary and final models. The sequence is shown in violet, together with the amino acids it codes for, in yellow. The preliminary model thus begins with the sequence GTCGCA. . . , which codes for the amino acids valine and alanine. In fact, though, almost all genes begin with the amino acid methionine, a feature of the final gene model. At the upper right, GRAIL displays the results of a database search for sequences similar to the final five-exon gene model. Close matches were found among species as diverse as soybean and the nematode *Caenorhabditis elegans*.

sequencing has not yet been achieved. But the potential benefits are great, and much of the instrumentation for sensitive detection of fluorescence signals has already proved useful for molecular sizing in mapping applications.

DEALING WITH THE DATA

Among the less visible challenges of the Human Genome Project is the daunting prospect of coping with all the data that success implies. Appropriate information systems are needed not only during data acquisition, but also for sophisticated data analysis and for the management and public distribution of unprecedented quantities of biological information. Further, because much of the challenge is interpreting genomic data and making the results available for scientific and technological applications, the challenge extends not just to the Human Genome Project, but also to the microbial genome program and to public- and private-sector programs focused on areas such as health effects, structural biology, and environmental remediation. Efforts in all these areas are the mandate of the DOE genome informatics program, whose products are already widely used in genome laboratories, general molecular biology and medical laboratories, biotechnology companies, and biopharmaceutical companies around the world.

The roles of laboratory data acquisition and management systems include the construction of genetic and physical maps, DNA sequencing, and gene expression analysis. These systems typically comprise databases for tracking biological materials and experimental procedures, software for controlling robots or other automated systems, and software for acquiring laboratory data and presenting it in useful form. Among such systems are physical mapping databases developed at Livermore and Los Alamos, robot control software developed at Berkeley and Livermore, and DNA sequence assembly software developed at the University of Arizona. These systems are the keys to efficient, cost-effective data production in both DOE laboratories and the many other laboratories that use them.

The interpretation of map and sequence data is the job of data analysis systems. These systems typically include task-specific computational engines, together with graphics and user-friendly interfaces that invite their use by biologists and other non–computer scientists. The genome informatics program is the world leader in developing automated systems for identifying genes in DNA sequence data from humans and other organisms, supporting efforts at Oak Ridge National Laboratory and elsewhere. The Oak Ridge–developed GRAIL system, illustrated in Figure 10, is a world-standard gene identification tool. In 1995 alone, more than 180 million base pairs of DNA were analyzed with GRAIL.

A third area of informatics reflects, in a sense, the ultimate product of the Human Genome Project—information readily available to the scientific and lay communities. Public resource databases must provide data and interpretive analyses to a worldwide research and development community. As this community of researchers expands and as the quantity of data grows, the challenges of maintaining accessible and useful databases likewise increase. For example, it is critical to develop scientific databases that "interoperate," sharing data and protocols so that users can expect answers to complex questions that demand information from geographically distributed data resources. As the genome project continues to provide data that interlink structural and functional biochemistry, molecular, cellular, and developmental biology, physiology and medicine, and environmental science, such interoperable databases will be the critical resources for both research and technology development. The DOE genome informatics program is crucial to the multiagency effort to develop just such databases. Systems now in place include the Genome Database of human genome map data at Johns Hopkins University, the Genome Sequence DataBase at the National Center for Genome Resources in Santa Fe, and the Molecular Structure Database at Brookhaven National Laboratory. ◆

Ethical, Legal, and Social Implications

AN ESSENTIAL DIMENSION OF GENOME RESEARCH

THE HUMAN GENOME PROJECT is rich with promise, but also fraught with social implications. We expect to learn the underlying causes of thousands of genetic diseases, including sickle cell anemia, Tay-Sachs disease, Huntington disease, myotonic dystrophy, cystic fibrosis, and many forms of cancer—and thus to predict the likelihood of their occurrence in any individual. Likewise, genetic information might be used to predict sensitivities to various industrial or environmental agents. The dangers of misuse and the potential threats to personal privacy are not to be taken lightly.

Both the DOE and the NIH devote a portion of their resources to studies of ethical, legal, and social implications.

In recognition of these important issues, both the DOE and the National Institutes of Health devote a portion of their resources to studies of the ethical, legal, and social implications (ELSI) of human genome research. Perhaps the most critical of social issues are the questions of privacy and fair use of genetic information. Most observers agree that personal knowledge of genetic susceptibility can be expected to serve us well, opening the door to more accurate diagnoses, preventive intervention, intensified screening, lifestyle changes, and early and effective treatment. But such knowledge has another side, too: the risk of anxiety, unwelcome changes in personal relationships, and the danger of stigmatization. Consider, for example, the impact of information that is likely to be incomplete and indeterminate (say, an indication of a 25 percent increase in the risk of cancer). And further, if handled carelessly, genetic information could threaten us with discrimination by potential employers and insurers. Other issues are perhaps less immediate than these personal concerns, but they are no less challenging. How, for example, are the "products" of the Human Genome Project to be patented and commercialized? How are the judicial, medical, and educational communities—not to mention the public at large—to be effectively educated about genetic research and its implications?

To confront all these issues, the NIH-DOE Joint Working Group on Ethical, Legal, and Social Implications of Human Genome Research was created in 1990 to coordinate ELSI policy and research between the two agencies. One focus of DOE activity has been to foster educational programs aimed both at private citizens and at policy-makers and educators. Fruits of these efforts include radio and television documentaries, high school curricula and other educational material, and science museum displays. In addition, the DOE has concentrated on issues associated with privacy and the confidentiality of genetic information, on workplace and commercialization issues (especially screening for susceptibilities to environmental or workplace agents), and on the implications of research findings regarding the interactions among multiple genes and environmental influences.

Whereas the issues raised by modern genome research are among the most challenging we face, they are not unprecedented. Issues of privacy, knotty questions of how knowledge is to be commercialized, problems of dealing with probabilistic risks, and the imperatives of education have all been confronted before. As usual, defensible perspectives and reasonable arguments, even precious rights, exist on opposing sides of every issue. It is a balance that must be sought. Accordingly, further study is needed, as well as continuing efforts to promote public awareness and understanding, as we strive to define policies for the intelligent use of the profound knowledge we seek about ourselves.

◆

THE AGE OF DISCOVERY was the age of da Gama, Columbus, and Magellan, an era when European civilization reached out to the Far East and thus filled many of the voids in its map of the world. But in a larger sense, we have never ceased from our exploration and discovery. Science has been unstinting over the ages in its efforts to complete our intellectual picture of the universe. In this century, our explorations have extended from the subatomic to the cosmic, as we have mapped the heavens to their farthest reaches and charted the properties of the most fleeting elementary particles. Nor have we neglected to look inward, seeking, as it were, to define the topography of the human body. Beginning with the first modern anatomical studies in the sixteenth century, we have added dramatically to our picture of human anatomy, physiology, and biochemistry. The Human Genome Project is thus the next stage in an epic voyage of discovery—a voyage that will bring us to a profound understanding of human biology.

In an important way, though, the genome project is very different from many of our exploratory adventures. It is spurred by a conviction of practical value, a certainty that human benefits will follow in the wake of success. The product of the Human Genome Project will be an enormously rich biological database, the key to tracking down every human gene—and thus to unveiling, and eventually to subverting, the causes of thousands of human diseases. The sequence of our genome will ultimately allow us to unlock the secrets of life's processes, the biochemical underpinnings of our senses and our memory, our development and our aging, our similarities and our differences.

It has further been said that the Human Genome Project is *guaranteed* to succeed: Its goal is nothing more assuming than a sequence of three billion characters. And we have a very good idea of how to read those characters. Unlike perilous voyages or searches for unknown subatomic particles, this venture is assured of its goal. But beyond a detailed picture of human DNA, no one can predict the form success will take. The genome project itself offers no promises of cancer cures or quick fixes for Alzheimer's disease, no detailed understanding of genius or schizophrenia. But if we are *ever* to uncover the mysteries of carcinogenesis, if we are *ever* to know how biochemistry contributes to mental illness and dementia, if we *ever* hope to really understand the processes of growth and development, we must first have a detailed map of the genetic landscape. That's what the Human Genome Project promises. In a way, it's a rather prosaic step, but what lies beyond is breathtaking.

The World Wide Web offers the easiest path to current news about the Human Genome Project. Good places to start include the following:

- DOE Human Genome Program—http://www.er.doe.gov/production/oher/hug_top.html
- NIH National Center for Human Genome Research—http://www.nchgr.nih.gov
- Human Genome Management Information System at Oak Ridge National Laboratory—http://www.ornl.gov/TechResources/Human_Genome/home.html
- Lawrence Berkeley National Laboratory Human Genome Center—http://www-hgc.lbl.gov/GenomeHome.html
- Lawrence Livermore National Laboratory Human Genome Center—http://www-bio.llnl.gov/bbrp/genome/genome.html
- Los Alamos National Laboratory Center for Human Genome Studies—http://www-ls.lanl.gov/LSwelcome.html
- The Genome Database at Johns Hopkins University School of Medicine—http://gdbwww.gdb.org/
- The National Center for Genome Resources—http://www.ncgr.org/

ACKNOWLEDGMENTS

This booklet was prepared at the request of the U.S. Department of Energy, Office of Health and Environmental Research, as an overview of the Human Genome Project, especially the role of the DOE in this international, multiagency effort. Though edited and produced at the Lawrence Berkeley National Laboratory, this account aims to reflect the full scope of the DOE-sponsored effort. In pursuit of this goal, the contributions of many have been essential. Within the Department of Energy, David A. Smith deserves special mention. He managed the DOE Human Genome Program until his retirement this year, and he was the principal catalyst of this effort to summarize its achievements. Also contributing program descriptions, illustrations, advice, and criticism: at DOE, Daniel W. Drell; at Berkeley, Michael Palazzolo, Christopher H. Martin, Sylvia Spengler, David Gilbert, Joseph M. Jaklevic, Eddy Rubin, Kerrie Whitelaw, and Manfred Zorn; at Lawrence Livermore National Laboratory, Anthony Carrano, Gregory G. Lennon, and Linda Ashworth; at Los Alamos National Laboratory, Robert K. Moyzis and Larry Deaven; at Oak Ridge National Laboratory, Lisa Stubbs; at the National Center for Genome Resources, Christopher Fields; and at Affymetrix, Robert J. Lipshutz. Behind the scenes, many others no doubt had a hand.

DOUGLAS VAUGHAN
Editor

Design: Debra Lamfers Design
Illustrations: Marilee Bailey

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098. PUB-773/July 1996.

Printed on recycled paper.

The National Human Genome Research Institute

A New Five-Year Plan for the U.S. Human Genome Program

Francis Collins and David Galas

Originally published in *Science* 262:43-46 (1993)

The U.S. Human Genome Project is part of an international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms. Thanks to advances in technology and a tightly focused effort, the project is on track with respect to its initial 5-year goals. Because 3 years have elapsed since these goals were set, and because a much more sophisticated and detailed understanding of what needs to be done and how to do it is now available, the goals have been refined and extended to cover the first 8 years (through September, 1998) of the 15 year genome initiative.

In 1990, the Human Genome Programs of the National Institutes of Health (NIH) and the Department of Energy (DOE) developed a joint research plan with specific goals for the first 5 years (FY 1991 - 1995) of the U.S. genome project. It has served as a valuable guide for both the research community and the agencies' administrative staff in developing and executing the genome project and assessing its progress for the past 3 years. Great strides have been made toward the achievement of the initial set of goals, particularly with respect to constructing detailed human genetic maps, improving physical maps of the human genome and the genomes of certain model organisms, developing improved technology for DNA sequencing and information handling, and defining the most urgent set of ethical, legal and social issues associated with the acquisition and use of large amounts of genetic information.

Progress toward achieving the first set of goals for the genome project appears to be on schedule, or in some instances, even ahead of schedule. Furthermore, technological improvements that could not have been anticipated in 1990 have in some areas changed the shape of the project and allowed more ambitious approaches. Earlier this year, it was therefore decided to update and extend the initial goals to address the scope of genome research beyond the completion of the original 5-year plan. A major purpose of revisiting the plan is to inform and provide a new guide to all participants in the genome project about the project's goals. To obtain the advice needed to develop the extended goals, NIH and DOE held a series of meetings with a large number of scientists and other interested scholars and representatives of the public, including many who previously had not been direct participants in the genome project. Reports of all these meetings are available from the Office of Communications of the National Human Genome Research Institute, and the Human Genome Management Information System of the DOE (2,3). Finally, a group of representative advisors from NIH and DOE drafted a set of new, extended goals for presentation to the National Advisory Council for Human Genome Research of the NIH and the Health and Environmental Research Advisory Committee of the DOE. These bodies have approved this document as a statement of their advice to the two agencies, and the following represents the goals for FY 1994 - 1998 (i.e. October 1, 1993 - September 30, 1998).

General Principles

Several general observations underlie the specific goals described here. The first observation is that successful development of new technology for genomic and genetic research has been essential to the achievements of the program to date and will continue to be critical in the future. It was clearly recognized, both in the 1988 NRC report (4) and in the first NIH-DOE plan, that attainment of the ambitious goals originally set for the genome project would require significant technological advances in all areas such as mapping, sequencing, informatics, and gene identification. As the genome project has proceeded, progress along a broad range of technological fronts has been conspicuous. Among the most notable of these developments have been (i) new types of genetic markers, such as microsatellites, that can be assayed by the polymerase chain reaction (PCR); (ii) improved vector systems for cloning large DNA fragments and better experimental strategies and computational methods for assembling those clones into large, overlapping sets (contigs) that compose useful physical maps; (iii) the definition of the sequence tagged site (STS) (5) as a common unit of physical mapping; and (iv) improved technology and automation for DNA sequencing. Further substantial improvements in technology are needed in all areas of genome research, especially in DNA sequencing, if the project is to stay on schedule and meet the demanding goals that are being set.

A second general observation concerns an evolution in the levels of biological organization at which genomic research will likely function over the next few years. Initially, attention was focused at the chromosome as the basic unit of genome analysis. Large-scale mapping efforts, in particular, were directed at construction of chromosome maps. The sophisticated genetic linkage maps now available and the detailed physical maps that are being produced are clear measures of the success of that approach. However, other units of study for the human genome project will also have increasing usefulness in the future. Therefore, further mapping efforts directed at both larger and smaller targets should be encouraged. At one end of the scale, "whole genome" mapping efforts, in which the entire genome is efficiently analyzed, have become feasible with developments in PCR application and robotics. These approaches generally produce relatively low resolution maps with current technology. At the other end of the scale, increasing attention needs to be paid to detailed mapping, sequencing and annotation of regions on the order of one to a few megabases in size. Although small in comparison to the whole genome, a megabase is still large in comparison to the capabilities of conventional molecular genetic analysis. Thus, development of efficient technology for approaching detailed analysis of several megabase sections of the genome will provide a useful bridge between conventional genetics and genomics, as well as a foundation for innovation from which future methods for analysis of larger regions may arise.

Third, a goal for identifying genes within maps and sequences, that was implicit in the original plan, has now been made explicit. The progress already made on the original goals, combined with promising new approaches to gene identification, allow this element of genome analysis to be given greater visibility. This increased emphasis on gene identification will greatly enrich the maps that are produced.

It must also be noted here, that, as in the original five-year plan, these goals again assume a funding level for the U.S. genome program of $200 million annually, adjusted for inflation. As the detailed cost analysis for the first five-year plan was performed in 1991, a cost of living increase must be added for all years beyond FY 1991. This funding level has not yet been achieved (see Table 1).

Table 1: Budget of the Human Genome Project for the NIH and the DOE (millions of dollars). (Note: Budgets for 1994 and 1995 have not yet been determined).

Fiscal Year	NIH	DOE	Total	1991 Projection of Needs
1991	87.4	47.4	134.8	135.1
1992	104.8	61.4	166.2	169.2
1993	106.1	64.5	170.6	218.9
1994	-	-	-	246.8
1995	-	-	-	259.9

International Aspects

The Human Genome Project is truly international in scope, as the original planners envisioned it. Its success to date has been possible because of major contributions from many countries and the extensive sharing of information and resources. It is hoped and anticipated that this spirit of international cooperation and sharing will continue. This coordination has been achieved largely by scientist to scientist interaction, facilitated by the Human Genome Organization (HUGO), which has taken on responsibility for some aspects of the management of the international chromosome workshops in particular. These workshops have served to encourage collaboration and the sharing of information and resources and to facilitate the expeditious completion of chromosome maps.

Several notable individual international collaborations have marked the genome project so far. One is the United States - United Kingdom collaboration on the sequencing of the *Caenorhabditis elegans* genome. Scientists at the Los Alamos National Laboratory are collaborating with Australian colleagues to develop a physical map of chromosome 16, and investigators at the Lawrence Livermore National Laboratory with Japanese scientists on a high resolution physical map of chromosome 21. Other joint efforts include the collaboration between the NIH and the Centre d'Etude du Polymorphism Humain (CEPH) on the genetic map of the human genome and the Whitehead/Massachusetts Institute of Technology-Genethon collaboration on the whole genome approach to the human physical map. These are but examples of the myriad interrelationships that have formed, generally spontaneously, among participating scientists.

Specific Goals

Genetic Map

The 2-5 cM human genetic map of highly informative markers called for in the original goals is expected to be completed on time. However, improvements to make the map more useful and accessible will still be needed. If the field develops as predicted, there will be an increasing demand for technology that allows the nonexpert to type families rapidly for medical research purposes. In addition, to study complex genetic diseases, there is a need to be able to easily test large numbers of individuals for many markers simultaneously. In the long run polymorphic markers that can be screened in a more automated fashion and methods of gene mapping that obviate the need for a standard set of polymorphic markers are also desirable.

Goals

- Complete the 2-5 cM map by 1995
- Develop technology for rapid genotyping
- Develop markers that are easier to use
- Develop new mapping technologies

Physical Map

An STS-based physical map of the human genome is expected to be available in the next 2-3 years, with some areas mapped in more detail than others and an average interval between markers of approximately 300 kilobases. However, such a map will not likely be sufficiently detailed to provide a substrate for sequencing or to be optimally useful to investigators searching for disease genes. The original goal of a physical map with STS markers at intervals of 100 kb remains realistic and useful and would serve both sequencers and mappers. Using widely available methods, a molecular biologist can isolate a gene that is within 100 kb of a mapped marker, and a sequencer can use such a map as the basis for preparing the DNA for sequencing. To the extent that they do not introduce statistical bias, the use of STS's with added value (such as those derived from polymorphic markers or genes) is encouraged, because such markers add to the usefulness of the map.

Goal

- Complete an STS map of the human genome at a resolution of 100 kb.

Physical maps of greater than 100 kb resolution are needed for DNA sequencing, for the purpose of finding genes and for other biological purposes. While a variety of options are being explored for creating such maps, the optimal approach is by no means clear. There is a need to develop new strategies for high resolution physical mapping as well as new cloning systems that are well integrated with advanced sequencing technology. Technology for sequencing is evolving rapidly. Therefore, preparation of sequence-ready sets of clones should be closely associated with an imminent intent to sequence.

There is a pressing need for clone libraries with improved stability and lower chimerism and other artifacts and a need for better technology for traveling from one STS to the next. A greater accessibility to clone libraries should also be encouraged.

DNA Sequencing

Although the goal of sequencing DNA at a cost of $0.50 per base pair may be met by 1996 as originally projected, the rate at which DNA can be sequenced will not be sufficient for sequencing the whole genome. Priority should be given during the next five years to increasing sequencing capacity by increasing the number of groups oriented toward large-scale production sequencing. Substantial new technology that will allow sequencing at higher rates and lower costs is also needed: both evolutionary technology developed from improvements in current gel-based approaches and revolutionary technology based on new principles. These developments will only occur if significantly greater financial resources can be invested in this area. It is estimated that an immediate investment of $100 million per year will be needed for sequencing technology alone, to allow the human genome to be sequenced by the year 2005.

Goals

- Develop efficient approaches to sequencing one- to several- megabase regions of DNA of high biological interest.
- Develop technology for high throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
- Build up sequencing capacity to a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

The standard model organisms should be sequenced as rapidly as possible, with *Escherichia coli* and *Saccharomyces cerevisiae* completed by 1998 or earlier and *C. elegans* nearing completion by 1998. It is often advantageous to sequence the corresponding regions of human and mouse DNA side-by-side in areas of high biological interest. The sequencing of full-length, mapped complementary DNA (cDNA) molecules is useful, especially if it is associated with technological innovation extensible to genomic sequencing.

The measurement of the cost of sequencing is complex and fraught with many uncertainties due to the diversity of approaches being used. However, we need to continue to reduce costs, as well as improve our ability to assess the accuracy of the sequence produced. This latter point must be addressed in future sequencing efforts. Cost will be highly dependent on the level of accuracy achieved.

Gene Identification

Identification of all the genes in the human genome and in the genomes of certain model organisms is an implicit part of the Human Genome Project. Although the previous 5-year plan did not explicitly identify this activity with a specific goal, progress in mapping and in technology now make it desirable to do so. With both genetic and physical maps of the human genome and the genomes of certain model organisms becoming available and large amounts of sequence data beginning to appear, it is important to develop better methods for identifying all the genes and incorporating all known genes onto the physical maps and the DNA sequences that are produced. This information will make the maps most useful to scientists studying the role of genes in health and disease. While many promising approaches are being explored, more development is needed in this area.

Goal

- Develop efficient methods of identifying genes and for placement of known genes on physical maps or sequenced DNA.

Technology Development

Development of new and improved technology is vital to the genome project. Certain technologies, such as automation and robotics, cut across many areas of genome research and need particular attention. Cooperation in technology development should be encouraged where possible, because it is likely to be more effective and efficient than competition and duplication. The technology developed must be expandable and exportable, the long term goal being to create technology that will be available in many basic science laboratories and allow the efficient sequencing of other genomes. Technology development is costly and has not been sufficiently funded.

Goal

- Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and to meet the needs of the Human Genome Project as a whole.

Model Organisms

Excellent progress has been made on the mouse genetic map, the *Drosophila* physical map, as well as the sequencing of the DNA of *E. coli*, *S. cerevisiae* and *C. elegans*. Many of the original goals for this area are likely to be exceeded. Completion of the mouse map and sequencing of all the selected model organism genomes continue to be high priorities. The current emphasis for sequencing of mouse DNA should be placed on sequencing of selected regions of high biologic interest side-by-side with the corresponding human DNA.

Goals

- Finish an STS map of the mouse at 300 Kb resolution
- Finish the sequence of the *E. coli* and *S. cerevisiae* genomes by 1998 or earlier
- Continue sequencing *C. elegans* and *Drosophila* genomes, with the aim of bringing *C. elegans* to near completion by 1998
- Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest

Informatics

In order to collect, organize and interpret the large amounts of complex mapping and sequencing data produced by the Human Genome Project, appropriate algorithms, software, database tools and operational infrastructure are required. The success of the genome project will depend, in large part, on the ease with which biologists can gain access to and use the information produced. Although considerable progress has been made in this area since the beginning of the genome project, there is a continuing need for improvements to stay current with evolving requirements. As the amount of information increases, the demand for it and the need for convenient access increase also. Thus, data management, data analysis and data distribution remain major goals for the future.

Goals

- Continue to create, develop and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among

databases
- Consolidate, distribute and continue to develop effective software for large-scale genome projects
- Continue to develop tools for comparing and interpreting genome information

Ethical, Legal and Social Implications (ELSI)

The ELSI components of the Human Genome programs of NIH and DOE are strongly connected with genomic research, so that policy discussions and the recommendations developed are couched in the reality of the science. To date, the focus of the ELSI programs has been on the most immediate potential applications in society of genome research. Four areas were identified by advisors to the ELSI program for initial emphasis: privacy of genetic information, safe and effective introduction of genetic information in the clinical setting, fairness in the use of genetic information and professional and public education. The program gives strong emphasis to understanding the ethnic, cultural, social and psychological influences that must inform policy development and service delivery. Initial policy options for genetic family studies, clinical genetic services, and health care coverage have been developed and reports on a range of urgent issues are expected by 1995.

As the genome project progresses, the need to prepare for broad public impact becomes increasingly important. Policies are needed to anticipate the potential consequences of widespread use of genetic tests for common conditions, such as genetic predisposition to certain cancers or genetic susceptibility to certain environmental agents. In addition, as the genetic elements of behavioral and other non-disease related traits are better understood, increased educational efforts will be needed to prevent stigmatization or discrimination based on these traits. Continued emphasis on public and professional education at all levels will be critical to achieving these goals. Mechanisms for developing policy options that build on the current research portfolio and actively involve the public, the relevant professions and the scientific community need to be developed.

Goals

- Continue to identify and define issues and develop policy options to address them
- Develop and disseminate policy options regarding genetic testing services with widespread potential use
- Foster greater acceptance of human genetic variation
- Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues

Training

There is a continuing need for individuals highly trained in the interdisciplinary sciences related to genome research. The original goal for supporting 600 trainees per year proved to be unattainable, because the capacity to train so many individuals in interdisciplinary sciences did not exist. However, now that a number of genome centers have been established, it is anticipated that training programs will expand. Although no numerical goal is specified, expansion of training activities should be encouraged, provided standards are kept high. Quality is more important than quantity.

Goal

- Continue to encourage training of scientists in interdisciplinary sciences related to genome research

Technology Transfer

Technology transfer is already occurring to a remarkable extent, as evidenced by the number of genome-related companies that are forming. Many interactions and collaborations have been established between genome researchers and the private sector. In addition to the need to transfer technology out of centers of genome research, there is also a need to increase the transfer of technology from other fields into the genome centers. Increased cooperation with industry, as well as continued cooperation between the agencies, is highly desirable. Care must be taken, however, to avoid conflicts of interest.

Goal

- Encourage and enhance technology transfer both into and out of centers of genome research

Outreach

It is essential to the success of the Human Genome Project that the products of genome research be made available to the community. However, only a subset of the total information is likely to be of interest at any one time, with the nature of the subset changing over time. Therefore, it is desirable to have flexible distribution systems that respond quickly to user demand. The private sector is best suited to this situation and has begun to play an active and highly valued role. This should be encouraged and facilitated where possible, including the provision of seed funding in some instances.

The NIH and DOE genome programs have adopted a rule for sharing of information: Newly developed data and materials are to be released within 6 months of their creation. This policy has been well accepted. In many instances, information has been released before the end of the six months.

Goals

- Cooperate with those who would set up distribution centers for genome materials.
- Share all information and materials within 6 months of their development. This should be accomplished by submission to public databases or repositories, or both, where appropriate.

Conclusion

To date the Human Genome Project has experienced gratifying success. However, enormous challenges remain. The technology that will allow the sequencing of the full human genome at reasonable cost must still be developed. Major support of research in this area is essential if the genome project is to succeed in the long run. The new goals described here are designed to address the long- and short-term needs of the project.

Although there is still debate about the need to sequence the entire genome, it is now more widely recognized that DNA sequence will reveal a wealth of biological information that could not be obtained in other ways. The sequence so far obtained from model organisms has demonstrated the existence of a large number of genes not previously suspected. For example, almost half the open reading frames identified in the genomic DNA of *C. elegans* appear to represent previously unidentified genes. Similar results have been observed in both *S. cerevisiae* and *E. coli* genomic DNA. Comparative sequence analysis has also confirmed the high degree of homology between genes across species. It is clear that sequence information represents a rich source for future investigation. Thus, the Human Genome Project must continue to pursue its ultimate goal, namely to obtain the complete human DNA sequence. At the same time, it is necessary to assure that technologies are developed that will allow the full interpretation of the DNA sequence once it is available. In order to increase emphasis on this area, an explicit goal related to gene identification has been added.

The genome project has already had a profound impact on biomedical research, as evidenced by the isolation of a number of genes associated with important diseases, such as Huntington's disease, amyotrophic lateral sclerosis, neurofibromatosis types 1 and 2, myotonic dystrophy, and fragile X syndrome. Genes that confer a predisposition to common diseases such as breast cancer, colon cancer, hypertension, diabetes and Alzheimer's disease have also been localized to specific chromosomal regions. All these discoveries benefitted from the information, resources and technologies developed by human genome research. As the genome project proceeds, many more exciting developments are expected including technology for studying the health effects of environmental agents, the ability to decipher the genomes of many other organisms, including countless microbes important to agriculture and the environment, as well as the identification of many more genes involved in disease. The technology and data produced by the genome project will provide a strong stimulus to broad areas of biological research and biotechnology. Exciting years lie ahead as the Human Genome Project moves toward its second set of 5-year goals.

References

1. U.S. Department of Health and Human Services, U.S. Department of Energy. *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.* (April, 1990).

2. National Institutes of Health/National Center for Human Genome Research, Office of Communications. Bethesda, MD 20892. Phone: 301/402-0911. Fax: 301/402-4570.

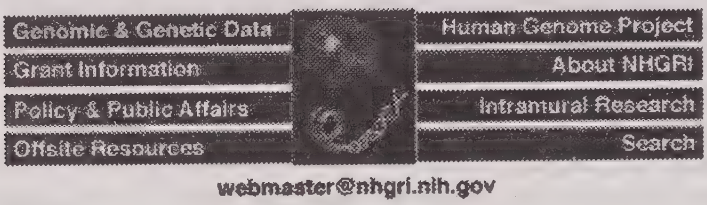
3. U.S. Department of Energy/Human Genome Management Information System. Oak Ridge National Laboratory, P.O. Box 20008, Oak Ridge, TN 37831-6050. Phone: 615-576-6669. Fax: 615/574-9188.

4. National Research Council, Committee on Mapping and Sequencing the Human Genome. *Mapping and Sequencing the Human Genome.* National Academy Press: Washington, D.C. (1988).

5. M.V. Olson, L. Hood, C. Cantor, and D. Botstein. "A common language for physical mapping of the human genome." *Science* 245:143-4 (1989).

Legend for Figure 1 (not shown)

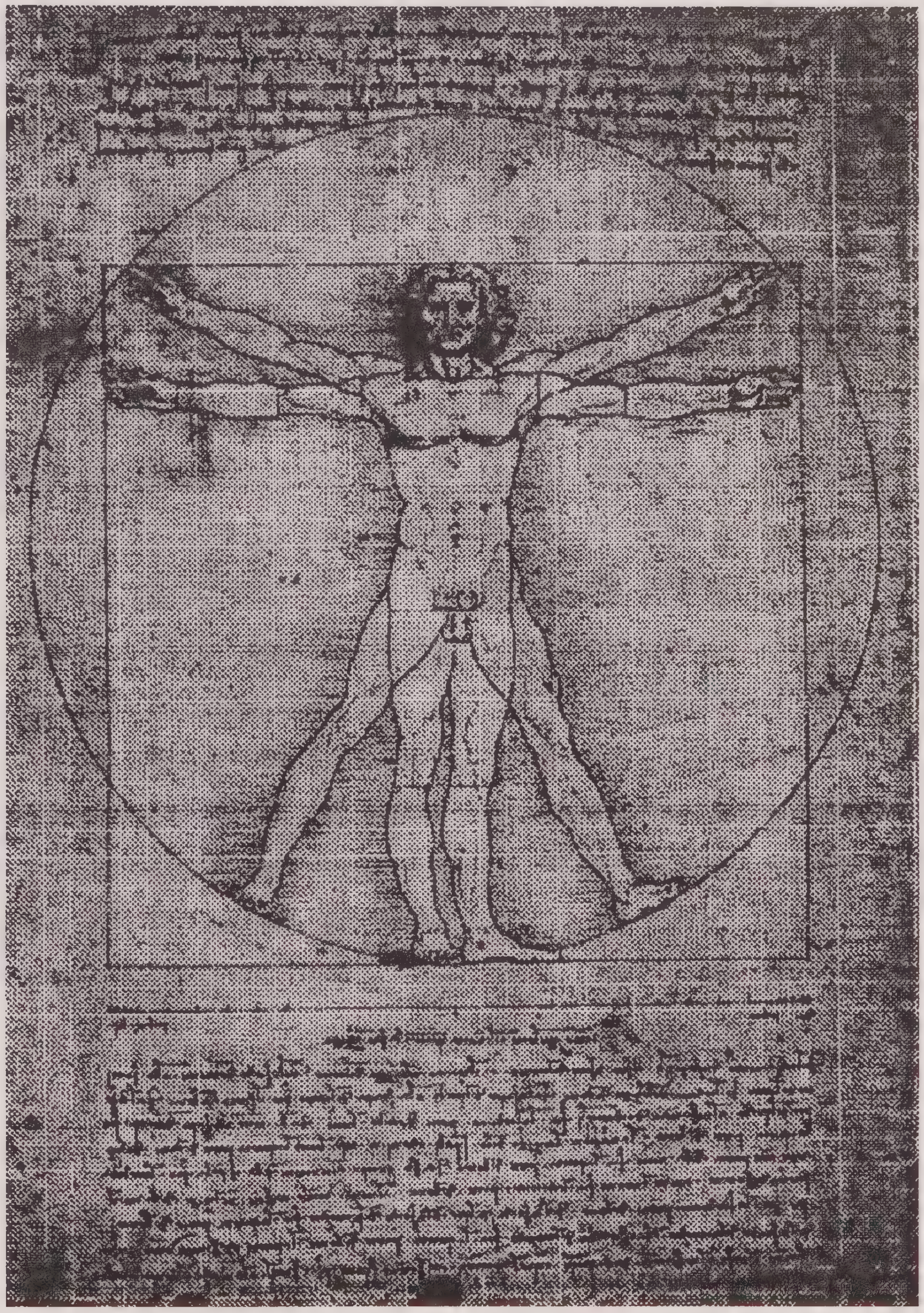
Graphic overview of the new goals for the human genome. A 2-5 centiMorgan genetic map is expected to be completed by 1995 and a physical map with STS markers every 100 kb by 1998. Efficient methods for gene identification need to be developed and refined. The DNA sequencing goal of 50 Megabases per year by 1998 includes all DNA, both human and model organisms, and assumes an exponential increase in sequencing capacity over time. Other important goals involving model organisms are not shown here, but are described in the text.


Genomic & Genetic Data
Grant Information
Policy & Public Affairs
Offsite Resources
Human Genome Project
About NHGRI
Intramural Research
Search
webmaster@nhgri.nih.gov

Genomic and Genetic Data | Grant Information | Policy and Public Affairs |
Offsite Resources | The Human Genome Project | About NHGRI | Intramural Research | Keyword Search

webmaster@nhgri.nih.gov

Primer on Molecular Genetics



DOE Human Genome Program

Primer on Molecular Genetics

Date Published: June 1992

U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research
Washington, DC 20585

The "Primer on Molecular Genetics" is taken from the June 1992 DOE *Human Genome 1991-92 Program Report*. The primer is intended to be an introduction to basic principles of molecular genetics pertaining to the genome project.

Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park
Oak Ridge, TN 37830

Voice: 423/576-6669
Fax: 423/574-9888
E-mail: bkq@ornl.gov

Contents

Primer on Molecular Genetics

Revised and expanded by Denise Casey (HGMIS) from the primer contributed by Charles Cantor and Sylvia Spengler (Lawrence Berkeley Laboratory) and published in the *Human Genome 1989–90 Program Report.*

Introduction 5
DNA 6
Genes 7
Chromosomes 8
Mapping and Sequencing the Human Genome 10
Mapping Strategies 11
  Genetic Linkage Maps 11
  Physical Maps 13
    Low-Resolution Physical Mapping 14
      Chromosomal map 14
      cDNA map 14
    High-Resolution Physical Mapping 14
      Macrorestriction maps: Top-down mapping 16
      Contig maps: Bottom-up mapping 17
Sequencing Technologies 18
  Current Sequencing Technologies 23
  Sequencing Technologies Under Development 24
  Partial Sequencing to Facilitate Mapping, Gene Identification 24
End Games: Completing Maps and Sequences; Finding Specific Genes 25
Model Organism Research 27
Informatics: Data Collection and Interpretation 27
Collecting and Storing Data 27
Interpreting Data 28
Mapping Databases 29
Sequence Databases 29
  Nucleic Acids (DNA and RNA) 29
  Proteins 30
Impact of the Human Genome Project 30
Glossary... 32

Introduction

The complete set of instructions for making an organism is called its genome. It contains the master blueprint for all cellular structures and activities for the lifetime of the cell or organism. Found in every nucleus of a person's many trillions of cells, the human genome consists of tightly coiled threads of deoxyribonucleic acid (DNA) and associated protein molecules, organized into structures called chromosomes (Fig. 1).

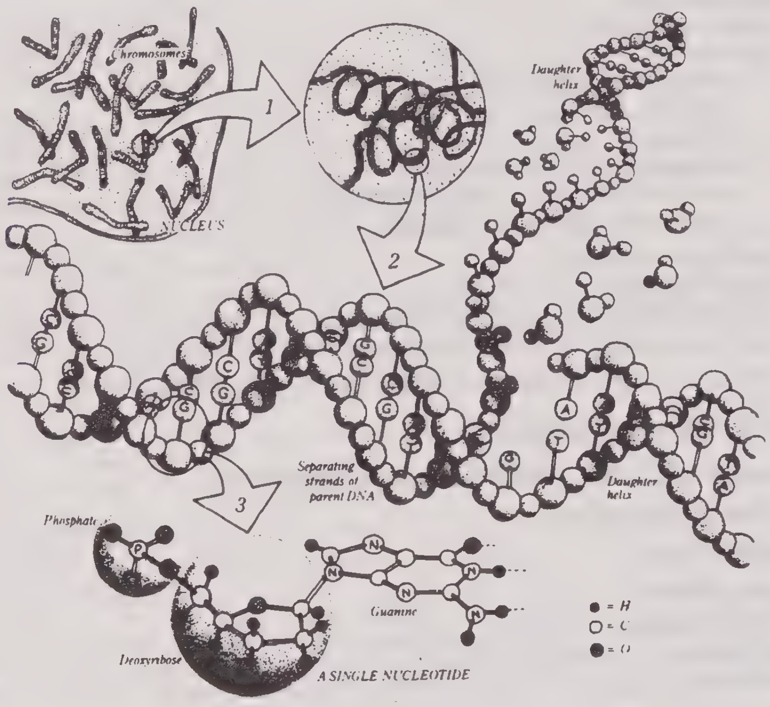


Fig. 1. The Human Genome at Four Levels of Detail. *Apart from reproductive cells (gametes) and mature red blood cells, every cell in the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA (1, 2). Each strand of DNA consists of repeating nucleotide units composed of a phosphate group, a sugar (deoxyribose), and a base (guanine, cytosine, thymine, or adenine) (3). Ordinarily, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between guanine and cytosine and between thymine and adenine. Each such linkage is a base pair (bp); some 3 billion bp constitute the human genome. The specificity of these base-pair linkages underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand; the nucleotide sequence (i.e., linear order of bases) of each strand is strictly determined. Each new double helix is a twin, an exact replica, of its parent. (Figure and caption text provided by the LBL Human Genome Center.)*

Primer on Molecular Genetics

If unwound and tied together, the strands of DNA would stretch more than 5 feet but would be only 50 trillionths of an inch wide. For each organism, the components of these slender threads encode all the information necessary for building and maintaining life, from simple bacteria to remarkably complex human beings. Understanding how DNA performs this function requires some knowledge of its structure and organization.

DNA

In humans, as in other higher organisms, a DNA molecule consists of two strands that wrap around each other to resemble a twisted ladder whose sides, made of sugar and phosphate molecules, are connected by "rungs" of nitrogen-containing chemicals called bases. Each strand is a linear arrangement of repeating similar units called nucleotides, which are each composed of one sugar, one phosphate, and a nitrogenous base (Fig. 2). Four different bases are present in DNA—adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar-phosphate backbone is called the DNA sequence; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits.

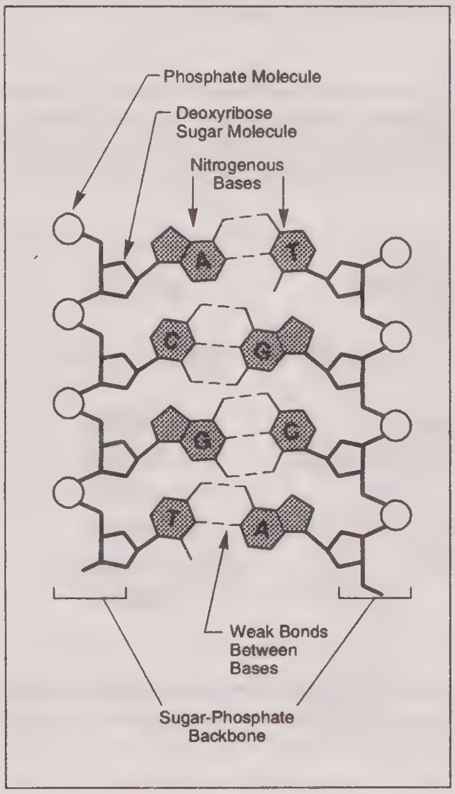


Fig. 2. DNA Structure. *The four nitrogenous bases of DNA are arranged along the sugar-phosphate backbone in a particular order (the DNA sequence), encoding all genetic instructions for an organism. Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). The two DNA strands are held together by weak bonds between the bases.*
A gene is a segment of a DNA molecule (ranging from fewer than 1 thousand bases to several million), located in a particular position on a specific chromosome, whose base sequence contains the information necessary for protein synthesis.

The two DNA strands are held together by weak bonds between the bases on each strand, forming base pairs (bp). Genome size is usually stated as the total number of base pairs; the human genome contains roughly 3 billion bp (Fig. 3).

Each time a cell divides into two daughter cells, its full genome is duplicated; for humans and other complex organisms, this duplication occurs in the nucleus. During cell division the DNA molecule unwinds and the weak bonds between the base pairs break, allowing the strands to separate. Each strand directs the synthesis of a complementary new strand, with free nucleotides matching up with their complementary bases on each of the separated strands. Strict base-pairing rules are adhered to—adenine will pair only with thymine (an A-T pair) and cytosine with guanine (a C-G pair). Each daughter cell receives one old and one new DNA strand (Figs. 1 and 4). The cell's adherence to these base-pairing rules ensures that the new strand is an exact copy of the old one. This minimizes the incidence of errors (mutations) that may greatly affect the resulting organism or its offspring.

Genes

Each DNA molecule contains many genes—the basic physical and functional units of heredity. A gene is a specific sequence of nucleotide bases, whose sequences carry the information required for constructing proteins, which provide the structural components of cells and tissues as well as enzymes for essential biochemical reactions. The human genome is estimated to comprise at least 100,000 genes.

Human genes vary widely in length, often extending over thousands of bases, but only about 10% of the genome is known to include the protein-coding sequences (exons) of genes. Interspersed within many genes are intron sequences, which have no coding function. The balance of the genome is thought to consist of other noncoding regions (such as control sequences and intergenic regions), whose functions are obscure. All living organisms are composed largely of proteins; humans can synthesize at least 100,000 different kinds. Proteins are large, complex molecules made up of long chains of subunits called amino acids. Twenty different kinds of amino acids are usually found in proteins. Within the gene, each specific sequence of three DNA bases (codons) directs the cell's protein-synthesizing machinery to add specific amino acids. For example, the base sequence ATG codes for the amino acid methionine. Since 3 bases code for 1 amino acid, the protein coded by an average-sized gene (3000 bp) will contain 1000 amino acids. The genetic code is thus a series of codons that specify which amino acids are required to make up specific proteins.

The protein-coding instructions from the genes are transmitted indirectly through messenger ribonucleic acid (mRNA), a transient intermediary molecule similar to a single strand of DNA. For the information within a gene to be expressed, a complementary RNA strand is produced (a process called transcription) from the DNA template in the nucleus. This

Comparative Sequence Sizes	Bases
• Largest known continuous DNA sequence (yeast chromosome 3)	350 Thousand
• *Escherichia coli* (bacterium) genome	4.6 Million
• Largest yeast chromosome now mapped	5.8 Million
• Entire yeast genome	15 Million
• Smallest human chromosome (Y)	50 Million
• Largest human chromosome (1)	250 Million
• Entire human genome	3 Billion

Fig. 3. Comparison of Largest Known DNA Sequence with Approximate Chromosome and Genome Sizes of Model Organisms and Humans. *A major focus of the Human Genome Project is the development of sequencing schemes that are faster and more economical.*

mRNA is moved from the nucleus to the cellular cytoplasm, where it serves as the template for protein synthesis. The cell's protein-synthesizing machinery then translates the codons into a string of amino acids that will constitute the protein molecule for which it codes (Fig. 5). In the laboratory, the mRNA molecule can be isolated and used as a template to synthesize a complementary DNA (cDNA) strand, which can then be used to locate the corresponding genes on a chromosome map. The utility of this strategy is described in the section on physical mapping.

Chromosomes

The 3 billion bp in the human genome are organized into 24 distinct, physically separate microscopic units called chromosomes. All genes are arranged linearly along the chromosomes. The nucleus of most human cells contains 2 sets of chromosomes, 1 set given by each parent. Each set has 23 single chromosomes—22 autosomes and an X or Y sex chromosome. (A normal female will have a pair of X chromosomes; a male will have an X

ORNL-DWG 91M-17361

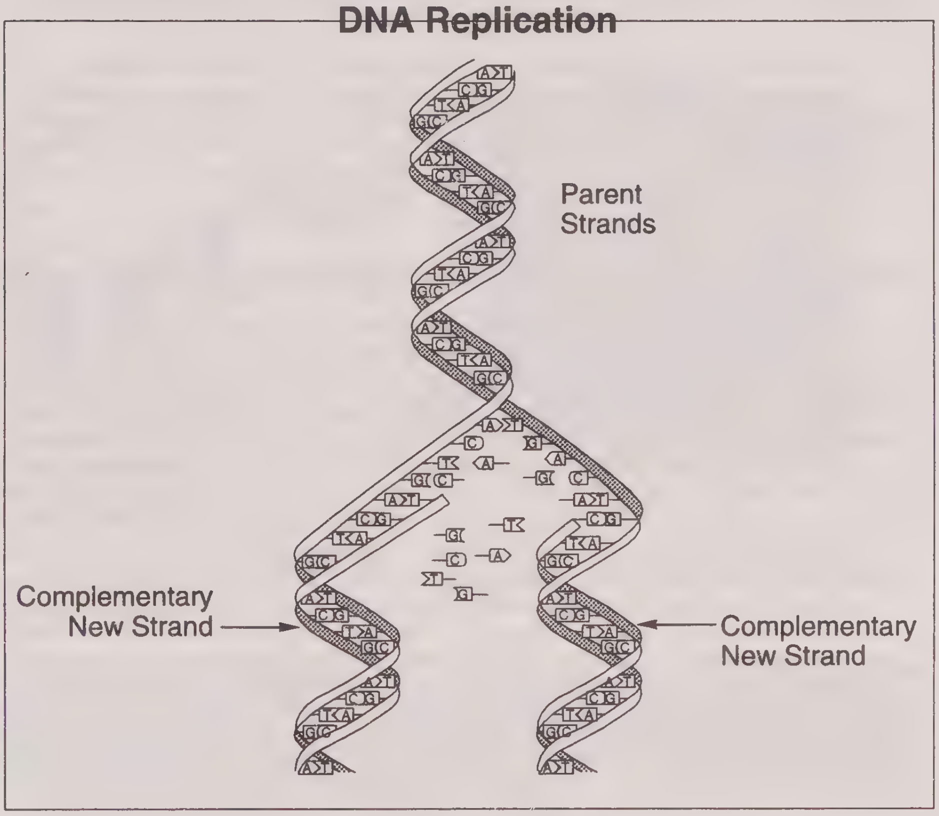


Fig. 4. DNA Replication. *During replication the DNA molecule unwinds, with each single strand becoming a template for synthesis of a new, complementary strand. Each daughter molecule, consisting of one old and one new DNA strand, is an exact copy of the parent molecule. [Source: adapted from* Mapping Our Genes—The Genome Projects: How Big, How Fast? *U.S. Congress, Office of Technology Assessment, OTA-BA-373 (Washington, D.C.: U.S. Government Printing Office, 1988).]*

and Y pair.) Chromosomes contain roughly equal parts of protein and DNA; chromosomal DNA contains an average of 150 million bases. DNA molecules are among the largest molecules now known.

Chromosomes can be seen under a light microscope and, when stained with certain dyes, reveal a pattern of light and dark bands reflecting regional variations in the amounts of A and T vs G and C. Differences in size and banding pattern allow the 24 chromosomes to be distinguished from each other, an analysis called a karyotype. A few types of major chromosomal abnormalities, including missing or extra copies of a chromosome or gross breaks and rejoinings (translocations), can be detected by microscopic examination; Down's syndrome, in which an individual's cells contain a third copy of chromosome 21, is diagnosed by karyotype analysis (Fig. 6). Most changes in DNA, however, are too subtle to be detected by this technique and require molecular analysis. These subtle DNA abnormalities (mutations) are responsible for many inherited diseases such as cystic fibrosis and sickle cell anemia or may predispose an individual to cancer, major psychiatric illnesses, and other complex diseases.

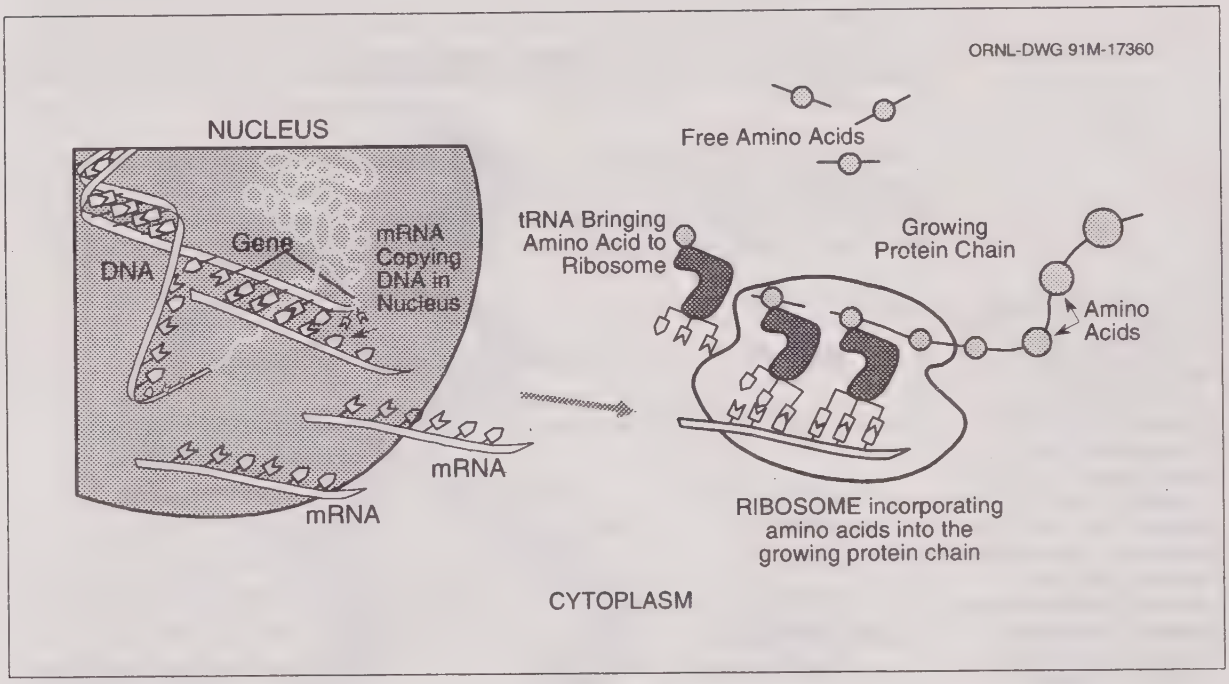

ORNL-DWG 91M-17360
NUCLEUS
Free Amino Acids
DNA
Gene
mRNA Copying DNA in Nucleus
tRNA Bringing Amino Acid to Ribosome
Growing Protein Chain
Amino Acids
mRNA
mRNA
RIBOSOME incorporating amino acids into the growing protein chain
CYTOPLASM

__Fig. 5. Gene Expression.__ When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replication. The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of bases (codons) forming the genetic code specify the particular amino acids that make up an individual protein. This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein. (Source: see Fig. 4.)

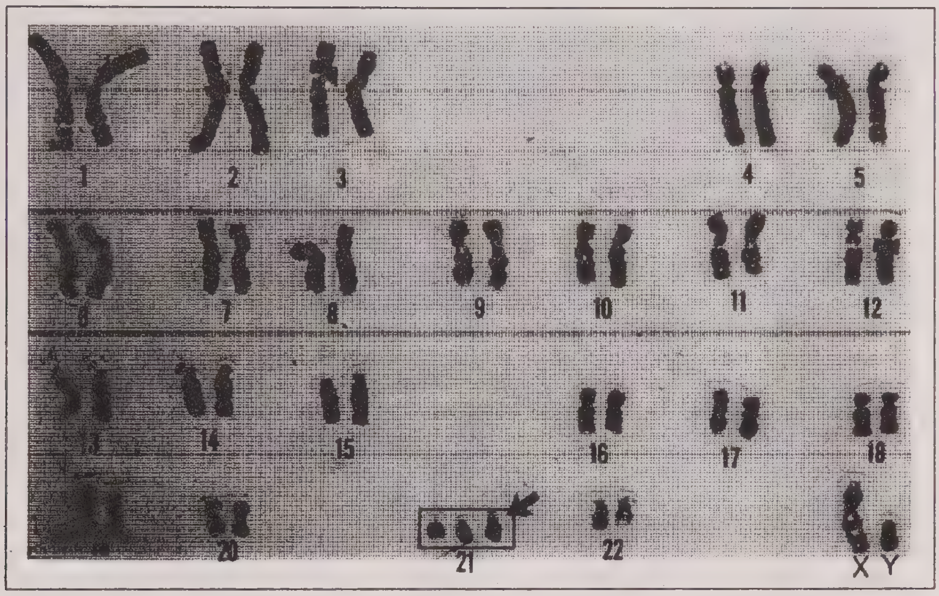


Fig. 6. ***Karyotype.*** *Microscopic examination of chromosome size and banding patterns allows medical laboratories to identify and arrange each of the 24 different chromosomes (22 pairs of autosomes and one pair of sex chromosomes) into a karyotype, which then serves as a tool in the diagnosis of genetic diseases. The extra copy of chromosome 21 in this karyotype identifies this individual as having Down's syndrome.*

Mapping and Sequencing the Human Genome

A primary goal of the Human Genome Project is to make a series of descriptive diagrams—maps—of each human chromosome at increasingly finer resolutions. Mapping involves (1) dividing the chromosomes into smaller fragments that can be propagated and char-acterized and (2) ordering (mapping) them to correspond to their respective locations on the chromosomes. After mapping is completed, the next step is to determine the base sequence of each of the ordered DNA fragments. The ultimate goal of genome research is to find all the genes in the DNA sequence and to develop tools for using this information in the study of human biology and medicine. Improving the instrumentation and techniques required for mapping and sequencing—a major focus of the genome project—will increase efficiency and cost-effectiveness. Goals include automating methods and optimizing techniques to extract the maximum useful information from maps and sequences.

A genome map describes the order of genes or other markers and the spacing between them on each chromosome. Human genome maps are constructed on several different scales or levels of resolution. At the coarsest resolution are genetic linkage maps, which depict the relative chromosomal locations of DNA markers (genes and other identifiable DNA sequences) by their patterns of inheritance. Physical maps describe the chemical characteristics of the DNA molecule itself.

Geneticists have already charted the approximate positions of over 2300 genes, and a start has been made in establishing high-resolution maps of the genome (Fig. 7). More-precise maps are needed to organize systematic sequencing efforts and plan new research directions.

Mapping Strategies

Genetic Linkage Maps

A genetic linkage map shows the relative locations of specific DNA markers along the chromosome. Any inherited physical or molecular characteristic that differs among individuals and is easily detectable in the laboratory is a potential genetic marker. Markers can be expressed DNA regions (genes) or DNA segments that have no known coding function but whose inheritance pattern can be followed. DNA sequence differences are especially useful markers because they are plentiful and easy to characterize precisely.

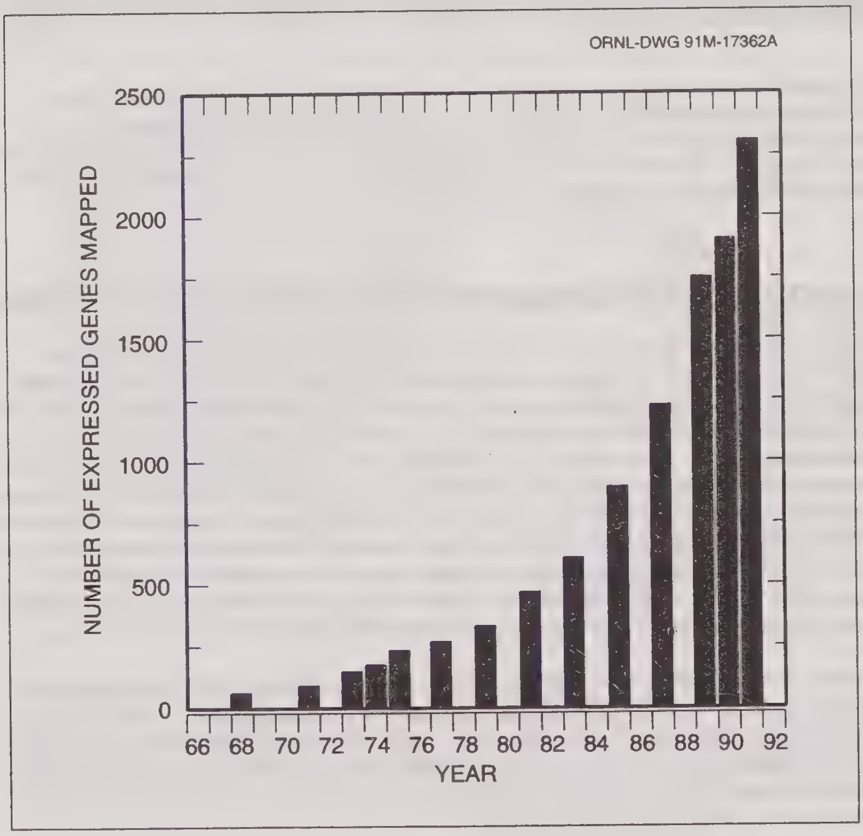


Fig. 7. Assignment of Genes to Specific Chromosomes. *The number of genes assigned (mapped) to specific chromosomes has greatly increased since the first autosomal (i.e., not on the X or Y chromosome) marker was mapped in 1968. Most of these genes have been mapped to specific bands on chromosomes. The acceleration of chromosome assignments is due to (1) a combination of improved and new techniques in chromosome sorting and band analysis, (2) data from family studies, and (3) the introduction of recombinant DNA technology. [Source: adapted from Victor A. McKusick, "Current Trends in Mapping Human Genes,"* The FASEB Journal *5(1), 12 (1991).]*

Markers must be polymorphic to be useful in mapping; that is, alternative forms must exist among individuals so that they are detectable among different members in family studies. Polymorphisms are variations in DNA sequence that occur on average once every 300 to 500 bp. Variations within exon sequences can lead to observable changes, such as differences in eye color, blood type, and disease susceptibility. Most variations occur within introns and have little or no effect on an organism's appearance or function, yet they are detectable at the DNA level and can be used as markers. Examples of these types of markers include (1) restriction fragment length polymorphisms (RFLPs), which reflect sequence variations in DNA sites that can be cleaved by DNA restriction enzymes (see box), and (2) variable number of tandem repeat sequences, which are short repeated sequences that vary in the number of repeated units and, therefore, in length (a characteristic easily measured). The human genetic linkage map is constructed by observing how frequently two markers are inherited together.

Two markers located near each other on the same chromosome will tend to be passed together from parent to child. During the normal production of sperm and egg cells, DNA strands occasionally break and rejoin in different places on the same chromosome or on the other copy of the same chromosome (i.e., the homologous chromosome). This process (called meiotic recombination) can result in the separation of two markers originally on the same chromosome (Fig. 8). The closer the markers are to each other—the more "tightly linked"—the less likely a recombination event will fall between and separate them. Recombination frequency thus provides an estimate of the distance between two markers.

On the genetic map, distances between markers are measured in terms of centimorgans (cM), named after the American geneticist Thomas Hunt Morgan. Two markers are said to be 1 cM apart if they are separated by recombination 1% of the time. A genetic distance of 1 cM is roughly equal to a physical distance of 1 million bp (1 Mb). The current resolution of most human genetic map regions is about 10 Mb.

The value of the genetic map is that an inherited disease can be located on the map by following the inheritance of a DNA marker present in affected individuals (but absent in unaffected individuals), even though the molecular basis of the disease may not yet be understood nor the responsible gene identified. Genetic maps have been used to find the exact chromosomal location of several important disease genes, including cystic fibrosis, sickle cell disease, Tay-Sachs disease, fragile X syndrome, and myotonic dystrophy.

HUMAN GENOME PROJECT GOALS

HUMAN GENOME PROJECT GOALS

	Resolution
• Complete a detailed human genetic map	2 Mb
• Complete a physical map	0.1 Mb
• Acquire the genome as clones	5 kb
• Determine the complete sequence	1 bp
• Find all the genes	

With the data generated by the project, investigators will determine the functions of the genes and develop tools for biological and medical applications.

One short-term goal of the genome project is to develop a high-resolution genetic map (2 to 5 cM); recent consensus maps of some chromosomes have averaged 7 to 10 cM between genetic markers. Genetic mapping resolution has been increased through the application of recombinant DNA technology, including in vitro radiation-induced chromosome fragmentation and cell fusions (joining human cells with those of other species to form hybrid cells) to create panels of cells with specific and varied human

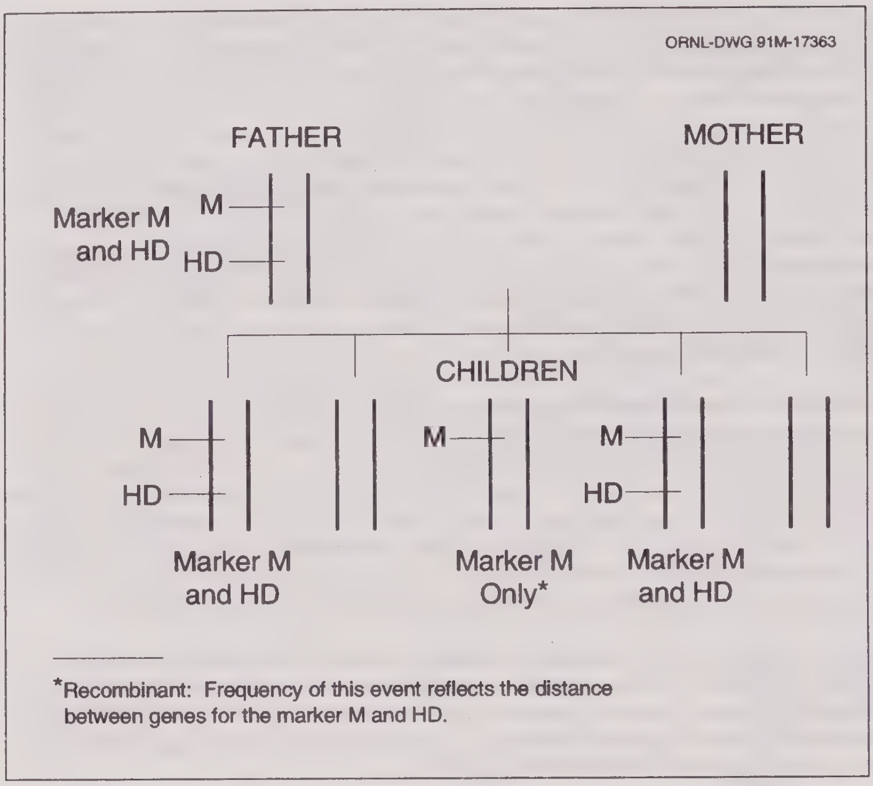


Fig. 8. Constructing a Genetic Linkage Map. *Genetic linkage maps of each chromosome are made by determining how frequently two markers are passed together from parent to child. Because genetic material is sometimes exchanged during the production of sperm and egg cells, groups of traits (or markers) originally together on one chromosome may not be inherited together. Closely linked markers are less likely to be separated by spontaneous chromosome rearrangements. In this diagram, the vertical lines represent chromosome 4 pairs for each individual in a family. The father has two traits that can be detected in any child who inherits them: a short known DNA sequence used as a genetic marker (M) and Huntington's disease (HD). The fact that one child received only a single trait (M) from that particular chromosome indicates that the father's genetic material recombined during the process of sperm production. The frequency of this event helps determine the distance between the two DNA sequences on a genetic map .*

chromosomal components. Assessing the frequency of marker sites remaining together after radiation-induced DNA fragmentation can establish the order and distance between the markers. Because only a single copy of a chromosome is required for analysis, even nonpolymorphic markers are useful in radiation hybrid mapping. [In meiotic mapping (described above), two copies of a chromosome must be distinguished from each other by polymorphic markers.]

Physical Maps

Different types of physical maps vary in their degree of resolution. The lowest-resolution physical map is the chromosomal (sometimes called cytogenetic) map, which is based on the distinctive banding patterns observed by light microscopy of stained chromosomes. A cDNA map shows the locations of expressed DNA regions (exons) on the chromosomal map. The more detailed cosmid contig map depicts the order of overlapping DNA fragments spanning the genome. A macrorestriction map describes the order and distance between enzyme cutting (cleavage) sites. The highest-resolution physical map is the complete elucidation of the DNA base-pair sequence of each chromosome in the human genome. Physical maps are described in greater detail below.

Low-Resolution Physical Mapping

Chromosomal map. In a chromosomal map, genes or other identifiable DNA fragments are assigned to their respective chromosomes, with distances measured in base pairs. These markers can be physically associated with particular bands (identified by cytogenetic staining) primarily by in situ hybridization, a technique that involves tagging the DNA marker with an observable label (e.g., one that fluoresces or is radioactive). The location of the labeled probe can be detected after it binds to its complementary DNA strand in an intact chromosome.

As with genetic linkage mapping, chromosomal mapping can be used to locate genetic markers defined by traits observable only in whole organisms. Because chromosomal maps are based on estimates of physical distance, they are considered to be physical maps. The number of base pairs within a band can only be estimated.

Until recently, even the best chromosomal maps could be used to locate a DNA fragment only to a region of about 10 Mb, the size of a typical band seen on a chromosome. Improvements in fluorescence in situ hybridization (FISH) methods allow orientation of DNA sequences that lie as close as 2 to 5 Mb. Modifications to in situ hybridization methods, using chromosomes at a stage in cell division (interphase) when they are less compact, increase map resolution to around 100,000 bp. Further banding refinement might allow chromosomal bands to be associated with specific amplified DNA fragments, an improvement that could be useful in analyzing observable physical traits associated with chromosomal abnormalities.

cDNA map. A cDNA map shows the positions of expressed DNA regions (exons) relative to particular chromosomal regions or bands. (Expressed DNA regions are those transcribed into mRNA.) cDNA is synthesized in the laboratory using the mRNA molecule as a template; base-pairing rules are followed (i.e., an A on the mRNA molecule will pair with a T on the new DNA strand). This cDNA can then be mapped to genomic regions.

Because they represent expressed genomic regions, cDNAs are thought to identify the parts of the genome with the most biological and medical significance. A cDNA map can provide the chromosomal location for genes whose functions are currently unknown. For disease-gene hunters, the map can also suggest a set of candidate genes to test when the approximate location of a disease gene has been mapped by genetic linkage techniques.

High-Resolution Physical Mapping

The two current approaches to high-resolution physical mapping are termed “top-down” (producing a macrorestriction map) and “bottom-up” (resulting in a contig map). With either strategy (described below) the maps represent ordered sets of DNA fragments that are generated by cutting genomic DNA with restriction enzymes (see Restriction Enzymes box at right). The fragments are then amplified by cloning or by polymerase chain reaction (PCR) methods (see DNA Amplification). Electrophoretic techniques are used to separate the fragments according to size into different bands, which can be visualized by direct DNA staining or by hybridization with DNA probes of interest. The use of purified chromosomes separated either by flow sorting from human cell lines or in hybrid cell lines allows a single chromosome to be mapped (see Separating Chromosomes box at right).

A number of strategies can be used to reconstruct the original order of the DNA fragments in the genome. Many approaches make use of the ability of single strands of DNA and/or RNA to hybridize—to form double-stranded segments by hydrogen bonding between complementary bases. The extent of sequence homology between the two strands can be

Restriction Enzymes: Microscopic Scalpels

Isolated from various bacteria, restriction enzymes recognize short DNA sequences and cut the DNA molecules at those specific sites. (A natural biological function of these enzymes is to protect bacteria by attacking viral and other foreign DNA.) Some restriction enzymes (rare-cutters) cut the DNA very infrequently, generating a small number of very large fragments (several thousand to a million bp). Most enzymes cut DNA more frequently, thus generating a large number of small fragments (less than a hundred to more than a thousand bp).

On average, restriction enzymes with

- 4-base recognition sites will yield pieces 256 bases long,
- 6-base recognition sites will yield pieces 4000 bases long, and
- 8-base recognition sites will yield pieces 64,000 bases long.

Since hundreds of different restriction enzymes have been characterized, DNA can be cut into many different small fragments.

Separating Chromosomes

Flow sorting

Pioneered at Los Alamos National Laboratory (LANL), flow sorting employs flow cytometry to separate, according to size, chromosomes isolated from cells during cell division when they are condensed and stable. As the chromosomes flow singly past a laser beam, they are differen-tiated by analyzing the amount of DNA present, and individual chromosomes are directed to specific collection tubes.

Somatic cell hybridization

In somatic cell hybridization, human cells and rodent tumor cells are fused (hybridized); over time, after the chromosomes mix, human chromosomes are preferentially lost from the hybrid cell until only one or a few remain. Those individual hybrid cells are then propagated and maintained as cell lines containing specific human chromosomes. Improvements to this technique have generated a number of hybrid cell lines, each with a specific single human chromosome.

inferred from the length of the double-stranded segment. Fingerprinting uses restriction map data to determine which fragments have a specific sequence (fingerprint) in common and therefore overlap. Another approach uses linking clones as probes for hybridization to chromosomal DNA cut with the same restriction enzyme.

Macrorestriction maps: Top-down mapping. In top-down mapping, a single chromosome is cut (with rare-cutter restriction enzymes) into large pieces, which are ordered and subdivided; the smaller pieces are then mapped further. The resulting macrorestriction maps depict the order of and distance between sites at which rare-cutter enzymes cleave (Fig. 9a). This approach yields maps with more continuity and fewer gaps between fragments than contig maps (see below), but map resolution is lower and may not be useful in finding particular genes; in addition, this strategy generally does not produce long stretches of mapped sites. Currently, this approach allows DNA pieces to be located in regions measuring about 100,000 bp to 1 Mb.

The development of pulsed-field gel (PFG) electrophoretic methods has improved the mapping and cloning of large DNA molecules. While conventional gel electrophoretic methods separate pieces less than 40 kb (1 kb = 1000 bases) in size, PFG separates molecules up to 10 Mb, allowing the application of both conventional and new mapping methods to larger genomic regions.

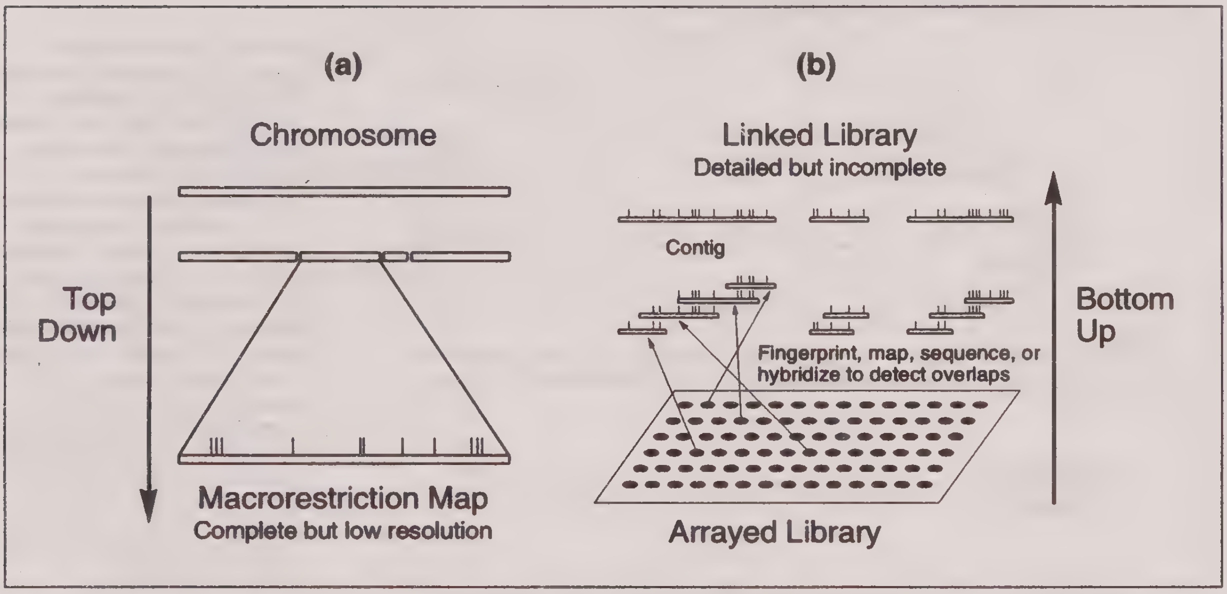

(a)
(b)
Chromosome
Linked Library
Detailed but incomplete
Contig
Top Down
Bottom Up
Fingerprint, map, sequence, or hybridize to detect overlaps
Macrorestriction Map
Complete but low resolution
Arrayed Library

*Fig. 9. **Physical Mapping Strategies.** Top-down physical mapping **(a)** produces maps with few gaps, but map resolution may not allow location of specific genes. Bottom-up strategies **(b)** generate extremely detailed maps of small areas but leave many gaps. A combination of both approaches is being used. [Source: Adapted from P. R. Billings et al., "New Techniques for Physical Mapping of the Human Genome,"* The FASEB Journal *5(1), 29 (1991).]*

Contig maps: Bottom-up mapping. The bottom-up approach involves cutting the chromosome into small pieces, each of which is cloned and ordered. The ordered fragments form contiguous DNA blocks (contigs). Currently, the resulting "library" of clones varies in size from 10,000 bp to 1 Mb (Fig. 9b). An advantage of this approach is the accessibility of these stable clones to other researchers. Contig construction can be verified by FISH, which localizes cosmids to specific regions within chromosomal bands.

Contig maps thus consist of a linked library of small overlapping clones representing a complete chromosomal segment. While useful for finding genes localized to a small area (under 2 Mb), contig maps are difficult to extend over large stretches of a chromosome because all regions are not clonable. DNA probe techniques can be used to fill in the gaps, but they are time consuming. Figure 10 is a diagram relating the different types of maps.

Technological improvements now make possible the cloning of large DNA pieces, using artificially constructed chromosome vectors that carry human DNA fragments as large as 1 Mb. These vectors are maintained in yeast cells as artificial chromosomes (YACs). (For more explanation, see DNA Amplification.) Before YACs were developed, the largest cloning vectors (cosmids) carried inserts of only 20 to 40 kb. YAC methodology drastically reduces the number of clones to be ordered; many YACs span entire human genes. A more detailed map of a large YAC insert can be produced by subcloning, a process in which fragments of the original insert are cloned into smaller-insert vectors. Because some YAC regions are unstable, large-capacity bacterial vectors (i.e., those that can accommodate large inserts) are also being developed.

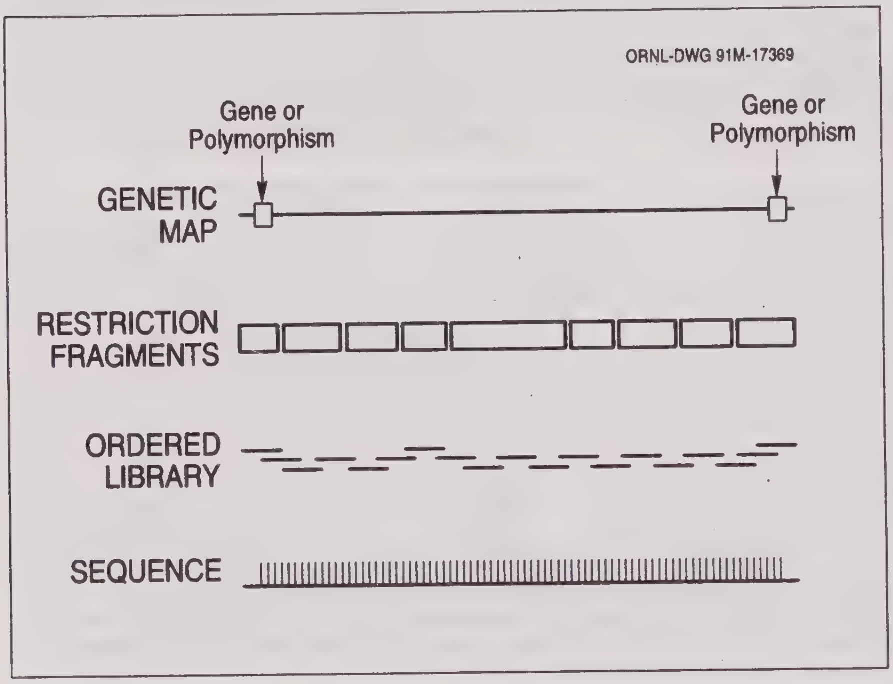


Fig. 10. Types of Genome Maps. *At the coarsest resolution, the genetic map measures recombination frequency between linked markers (genes or polymorphisms). At the next resolution level, restriction fragments of 1 to 2 Mb can be separated and mapped. Ordered libraries of cosmids and YACs have insert sizes from 40 to 400 kb. The base sequence is the ultimate physical map. Chromosomal mapping (not shown) locates genetic sites in relation to bands on chromosomes (estimated resolution of 5 Mb); new in situ hybridization techniques can place loci 100 kb apart. These direct strategies link the other four mapping approaches diagramed here. [Source: see Fig. 9.]*

Sequencing Technologies

The ultimate physical map of the human genome is the complete DNA sequence—the determination of all base pairs on each chromosome. The completed map will provide biologists with a Rosetta stone for studying human biology and enable medical researchers to begin to unravel the mechanisms of inherited diseases. Much effort continues to be spent locating genes; if the full sequence were known, emphasis could shift to determining gene function. The Human Genome Project is creating research tools for 21st-century biology, when the goal will be to understand the sequence and functions of the genes residing therein.

Achieving the goals of the Human Genome Project will require substantial improvements in the rate, efficiency, and reliability of standard sequencing procedures. While technological advances are leading to the automation of standard DNA purification, separation, and detection steps, efforts are also focusing on the development of entirely new sequencing methods that may eliminate some of these steps. Sequencing procedures currently involve first subcloning DNA fragments from a cosmid or bacteriophage library into special sequencing vectors that carry shorter pieces of the original cosmid fragments (Fig. 11). The next step is to make the subcloned fragments into sets of nested fragments differing in length by one nucleotide, so that the specific base at the end of each successive fragment is detectable after the fragments have been separated by gel electrophoresis. Current sequencing technologies are discussed later.

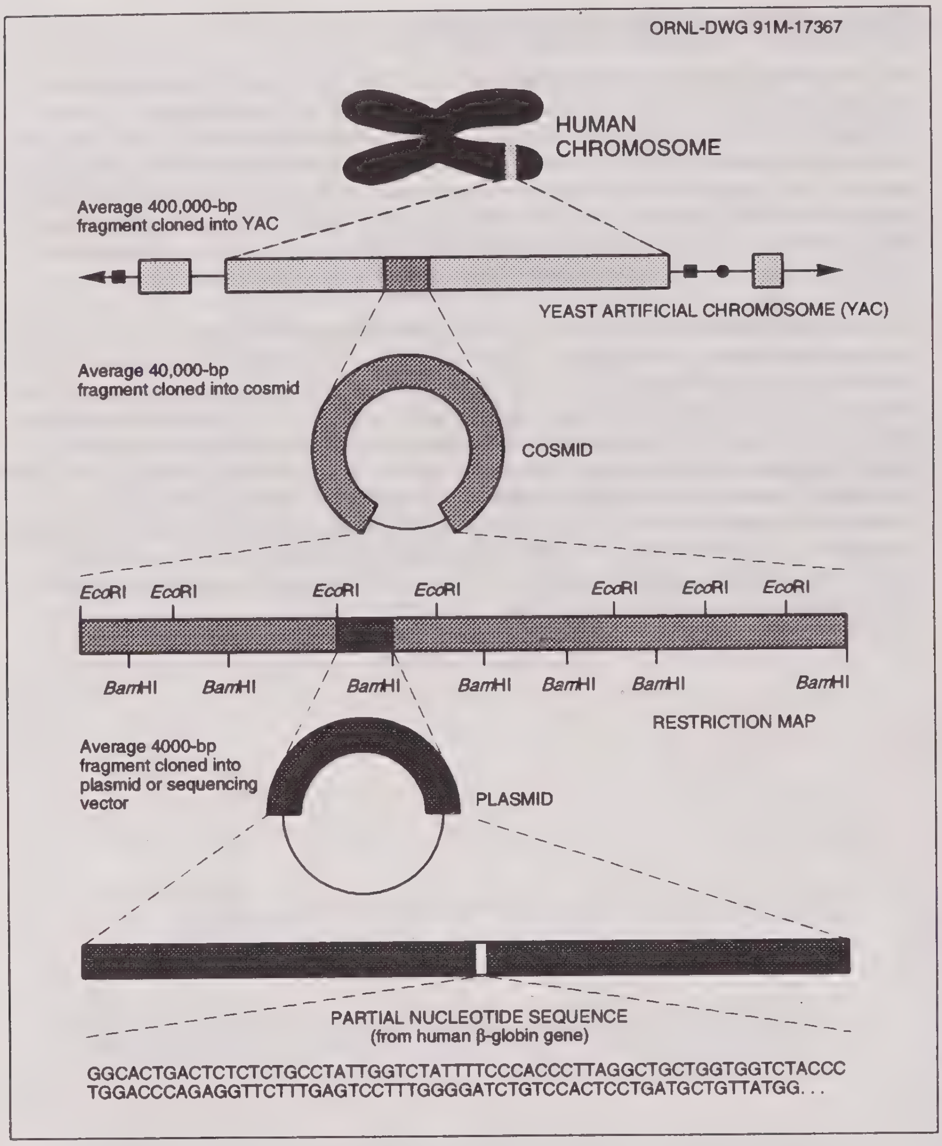

ORNL-DWG 91M-17367
HUMAN CHROMOSOME
Average 400,000-bp fragment cloned into YAC
YEAST ARTIFICIAL CHROMOSOME (YAC)
Average 40,000-bp fragment cloned into cosmid
COSMID
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
BamHI
BamHI
BamHI
BamHI
BamHI
BamHI
BamHI
RESTRICTION MAP
Average 4000-bp fragment cloned into plasmid or sequencing vector
PLASMID
PARTIAL NUCLEOTIDE SEQUENCE
(from human β-globin gene)
GGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCC
TGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGG. . .

Fig. 11. Constructing Clones for Sequencing. *Cloned DNA molecules must be made progressively smaller and the fragments subcloned into new vectors to obtain fragments small enough for use with current sequencing technology. Sequencing results are compiled to provide longer stretches of sequence across a chromosome. (Source: adapted from David A. Micklos and Greg A. Freyer,* DNA Science, A First Course in Recombinant DNA Technology, *Burlington, N.C.: Carolina Biological Supply Company, 1990.)*

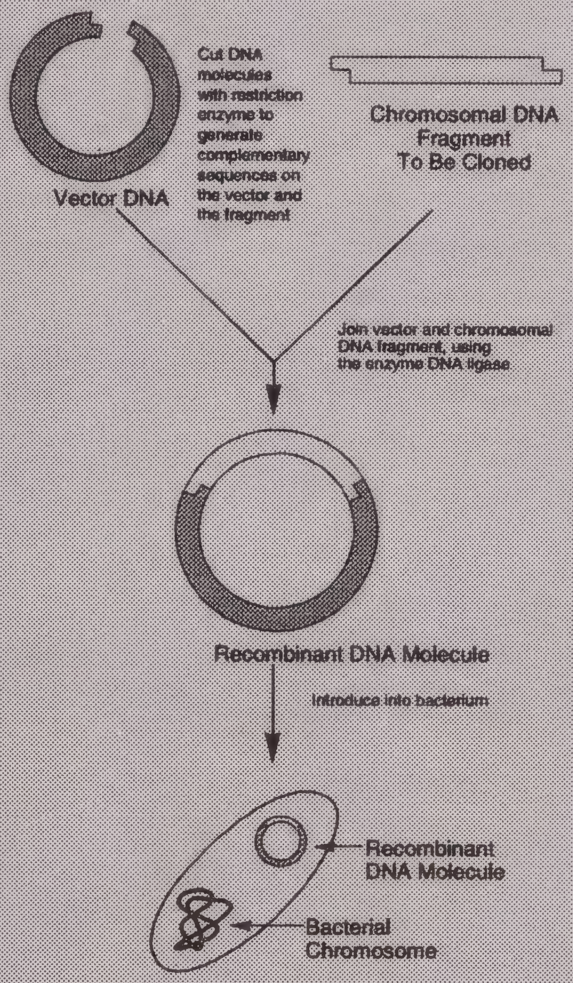
DNA Amplification: Cloning and Polymerase Chain Reaction (PCR)

Cloning (in vivo DNA amplification)

Cloning involves the use of recombinant DNA technology to propagate DNA fragments inside a foreign host. The fragments are usually isolated from chromosomes using restriction enzymes and then united with a carrier (a vector). Following introduction into suitable host cells, the DNA fragments can then be reproduced along with the host cell DNA. Vectors are DNA molecules originating from viruses, bacteria, and yeast cells. They accommodate various sizes of foreign DNA fragments ranging from 12,000 bp for bacterial vectors (plasmids and cosmids) to 1 Mb for yeast vectors (yeast artificial chromosomes). Bacteria are most often the hosts for these inserts, but yeast and mammalian cells are also used **(a)**.

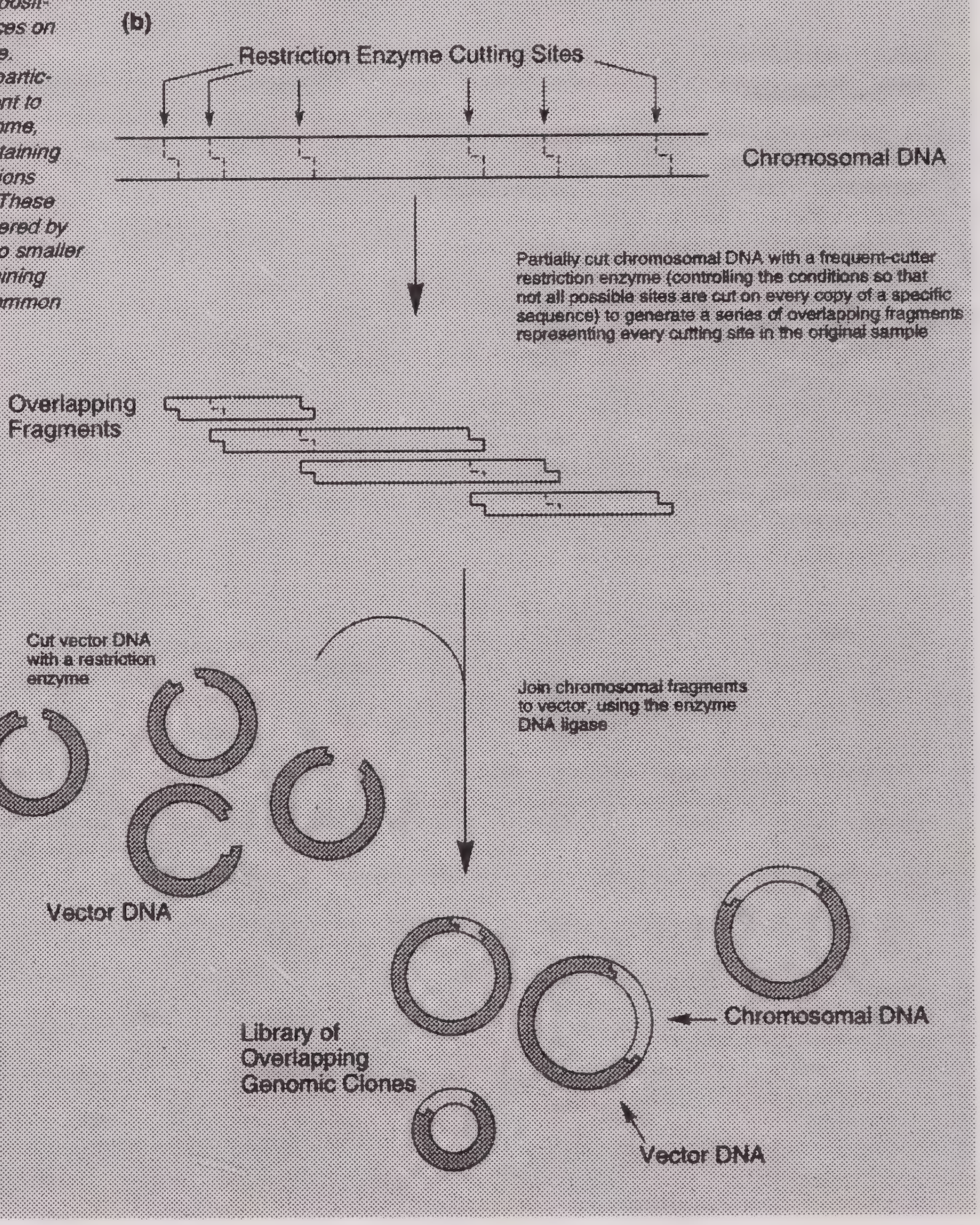
Cloning procedures provide unlimited material for experimental study. A random (unordered) set of cloned DNA fragments is called a library. Genomic libraries are sets of overlapping fragments encompassing an entire genome **(b)**. Also available are chromosome-specific libraries, which consist of fragments derived from source DNA enriched for a particular chromosome. (See Separating Chromosomes box.)

(a)


Cut DNA molecules with restriction enzyme to generate complementary sequences on the vector and the fragment
Vector DNA
Chromosomal DNA Fragment To Be Cloned
Join vector and chromosomal DNA fragment, using the enzyme DNA ligase
Recombinant DNA Molecule
Introduce into bacterium
Recombinant DNA Molecule
Bacterial Chromosome

(a) ***Cloning DNA in Plasmids.*** *By fragmenting DNA of any origin (human, animal, or plant) and inserting it in the DNA of rapidly reproducing foreign cells, billions of copies of a single gene or DNA segment can be produced in a very short time. DNA to be cloned is inserted into a plasmid (a small, self-replicating circular molecule of DNA) that is separate from chromosomal DNA. When the recombinant plasmid is introduced into bacteria, the newly inserted segment will be replicated along with the rest of the plasmid.*

(b) ***Constructing an Overlapping Clone Library.*** *A collection of clones of chromosomal DNA, called a library, has no obvious order indicating the original positions of the cloned pieces on the uncut chromosome. To establish that two particular clones are adjacent to each other in the genome, libraries of clones containing partly overlapping regions must be constructed. These clone libraries are ordered by dividing the inserts into smaller fragments and determining which clones share common DNA sequences.*


(b)
Restriction Enzyme Cutting Sites
Chromosomal DNA
Overlapping Fragments
Cut vector DNA with a restriction enzyme
Vector DNA
Library of Overlapping Genomic Clones
Chromosomal DNA
Vector DNA

Partially cut chromosomal DNA with a frequent-cutter restriction enzyme (controlling the conditions so that not all possible sites are cut on every copy of a specific sequence) to generate a series of overlapping fragments representing every cutting site in the original sample

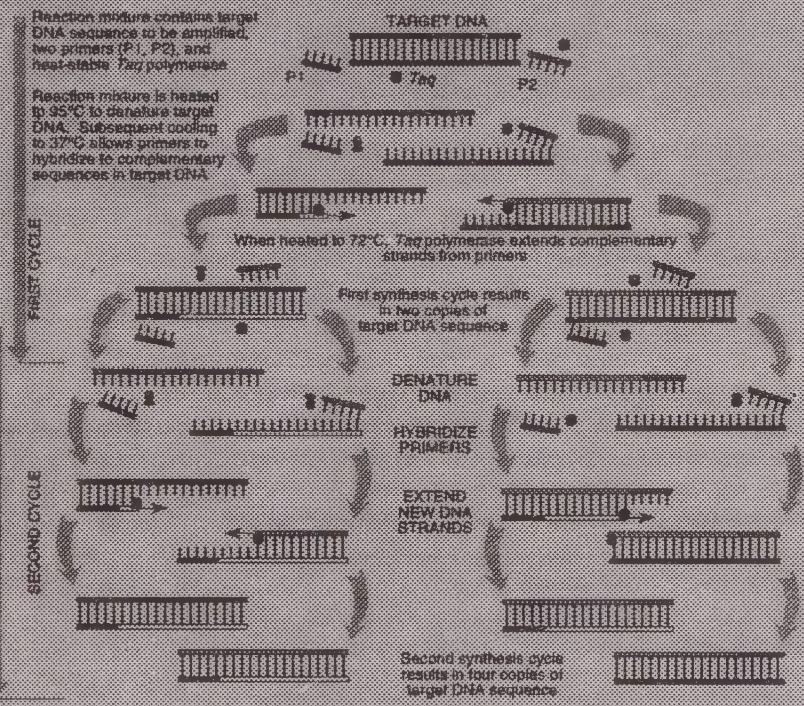
Join chromosomal fragments to vector, using the enzyme DNA ligase

PCR (in vitro DNA amplification)

Described as being to genes what Gutenberg's printing press was to the written word, PCR can amplify a desired DNA sequence of any origin (virus, bacteria, plant, or human) hundreds of millions of times in a matter of hours, a task that would have required several days with recombinant technology. PCR is especially valuable because the reaction is highly specific, easily automated, and capable of amplifying minute amounts of sample. For these reasons, PCR has also had a major impact on clinical medicine, genetic disease diagnostics, forensic science, and evolutionary biology.

PCR is a process based on a specialized polymerase enzyme, which can synthesize a complementary strand to a given DNA strand in a mixture containing the 4 DNA bases and 2 DNA fragments (primers, each about 20 bases long) flanking the target sequence. The mixture is heated to separate the strands of double-stranded DNA containing the target sequence and then cooled to allow (1) the primers to find and bind to their complementary sequences on the separated strands and (2) the polymerase to extend the primers into new complementary strands. Repeated heating and cooling cycles multiply the target DNA exponentially, since each new double strand separates to become two templates for further synthesis. In about 1 hour, 20 PCR cycles can amplify the target by a millionfold.

DNA Amplification Using PCR


Reaction mixture contains target DNA sequence to be amplified, two primers (P1, P2), and heat-stable *Taq* polymerase
TARGET DNA
P1
Taq
P2
Reaction mixture is heated to 95°C to denature target DNA. Subsequent cooling to 37°C allows primers to hybridize to complementary sequences in target DNA
When heated to 72°C, *Taq* polymerase extends complementary strands from primers
FIRST CYCLE
First synthesis cycle results in two copies of target DNA sequence
DENATURE DNA
HYBRIDIZE PRIMERS
EXTEND NEW DNA STRANDS
SECOND CYCLE
Second synthesis cycle results in four copies of target DNA sequence

Source: *DNA Science*, see Fig. 11.

Current Sequencing Technologies

The two basic sequencing approaches, Maxam-Gilbert and Sanger, differ primarily in the way the nested DNA fragments are produced. Both methods work because gel electrophoresis produces very high resolution separations of DNA molecules; even fragments that differ in size by only a single nucleotide can be resolved. Almost all steps in these sequencing methods are now automated. Maxam-Gilbert sequencing (also called the chemical degradation method) uses chemicals to cleave DNA at specific bases, resulting in fragments of different lengths. A refinement to the Maxam-Gilbert method known as multiplex sequencing enables investigators to analyze about 40 clones on a single DNA sequencing gel. Sanger sequencing (also called the chain termination or dideoxy method) involves using an enzymatic procedure to synthesize DNA chains of varying length in four different reactions, stopping the DNA replication at positions occupied by one of the four bases, and then determining the resulting fragment lengths (Fig. 12).

These first-generation gel-based sequencing technologies are now being used to sequence small regions of interest in the human genome. Although investigators could use existing technology to sequence whole chromosomes, time and cost considerations make large-scale sequencing projects of this nature impractical. The smallest human chromosome (Y) contains 50 Mb; the largest (chromosome 1) has 250 Mb. The largest continuous DNA sequence obtained thus far, however, is approximately 350,000 bp, and the best available equipment can sequence only 50,000 to 100,000 bases per year at an approximate cost of $1 to $2 per base. At that rate, an unacceptable 30,000 work-years and at least $3 billion would be required for sequencing alone.

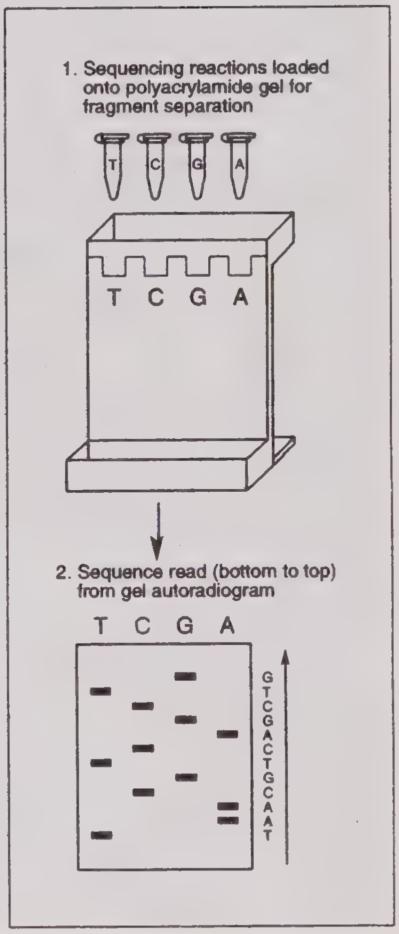

1. Sequencing reactions loaded onto polyacrylamide gel for fragment separation
T
C
G
A
T C G A
2. Sequence read (bottom to top) from gel autoradiogram
T C G A
GTCGACTGCAAT

Fig. 12. DNA Sequencing. *Dideoxy sequencing (also called chain-termination or Sanger method) uses an enzymatic procedure to synthesize DNA chains of varying lengths, stopping DNA replication at one of the four bases and then determining the resulting fragment lengths. Each sequencing reaction tube (T, C, G, and A) in the diagram contains*

- *a DNA template, a primer sequence, and a DNA polymerase to initiate synthesis of a new strand of DNA at the point where the primer is hybridized to the template;*
- *the four deoxynucleotide triphosphates (dATP, dTTP, dCTP, and dGTP) to extend the DNA strand;*
- *one labeled deoxynucleotide triphosphate (using a radioactive element or dye); and*
- *one* di*deoxynucleotide triphosphate, which terminates the growing chain wherever it is incorporated. Tube A has* di*dATP, tube C has* di*dCTP, etc.*

For example, in the A reaction tube the ratio of the dATP to di*dATP is adjusted so that each tube will have a collection of DNA fragments with a* di*dATP incorporated for each adenine position on the template DNA fragments. The fragments of varying length are then separated by electrophoresis (1) and the positions of the nucleotides analyzed to determine sequence. The fragments are separated on the basis of size, with the shorter fragments moving faster and appearing at the bottom of the gel. Sequence is read from bottom to top (2). (Source: see Fig. 11.)*

Sequencing Technologies Under Development

A major focus of the Human Genome Project is the development of automated sequencing technology that can accurately sequence 100,000 or more bases per day at a cost of less than $.50 per base. Specific goals include the development of sequencing and detection schemes that are faster and more sensitive, accurate, and economical. Many novel sequencing technologies are now being explored, and the most promising ones will eventually be optimized for widespread use.

Second-generation (interim) sequencing technologies will enable speed and accuracy to increase by an order of magnitude (i.e., 10 times greater) while lowering the cost per base. Some important disease genes will be sequenced with such technologies as (1) high-voltage capillary and ultrathin electrophoresis to increase fragment separation rate and (2) use of resonance ionization spectroscopy to detect stable isotope labels.

Third-generation gel-less sequencing technologies, which aim to increase efficiency by several orders of magnitude, are expected to be used for sequencing most of the human genome. These developing technologies include (1) enhanced fluorescence detection of individual labeled bases in flow cytometry, (2) direct reading of the base sequence on a DNA strand with the use of scanning tunneling or atomic force microscopies, (3) enhanced mass spectrometric analysis of DNA sequence, and (4) sequencing by hybridization to short panels of nucleotides of known sequence. Pilot large-scale sequencing projects will provide opportunities to improve current technologies and will reveal challenges investigators may encounter in larger-scale efforts.

Partial Sequencing To Facilitate Mapping, Gene Identification

Correlating mapping data from different laboratories has been a problem because of differences in generating, isolating, and mapping DNA fragments. A common reference system designed to meet these challenges uses partially sequenced unique regions (200 to 500 bp) to identify clones, contigs, and long stretches of sequence. Called sequence tagged sites (STSs), these short sequences have become standard markers for physical mapping.

Because coding sequences of genes represent most of the potentially useful information content of the genome (but are only a fraction of the total DNA), some investigators have begun partial sequencing of cDNAs instead of random genomic DNA. (cDNAs are derived from mRNA sequences, which are the transcription products of expressed genes.) In addition to providing unique markers, these partial sequences [termed expressed sequence tags (ESTs)] also identify expressed genes. This strategy can thus provide a means of rapidly identifying most human genes. Other applications of the EST approach include determining locations of genes along chromosomes and identifying coding regions in genomic sequences.

End Games: Completing Maps and Sequences; Finding Specific Genes

Starting maps and sequences is relatively simple; finishing them will require new strategies or a combination of existing methods. After a sequence is determined using the methods described above, the task remains to fill in the many large gaps left by current mapping methods. One approach is single-chromosome microdissection, in which a piece is physically cut from a chromosomal region of particular interest, broken up into smaller pieces, and amplified by PCR or cloning (see DNA Amplification). These fragments can then be mapped and sequenced by the methods previously described.

Chromosome walking, one strategy for filling in gaps, involves hybridizing a primer of known sequence to a clone from an unordered genomic library and synthesizing a short complementary strand (called "walking" along a chromosome). The complementary strand is then sequenced and its end used as the next primer for further walking; in this way the adjacent, previously unknown, region is identified and sequenced. The chromosome is thus systematically sequenced from one end to the other. Because primers must be synthesized chemically, a disadvantage of this technique is the large number of different primers needed to walk a long distance. Chromosome walking is also used to locate specific genes by sequencing the chromosomal segments between markers that flank the gene of interest (Fig. 13).

The current human genetic map has about 1000 markers, or 1 marker spaced every 3 million bp; an estimated 100 genes lie between each pair of markers. Higher-resolution genetic maps have been made in regions of particular interest. New genes can be located by combining genetic and physical map information for a region. The genetic map basically describes gene order. Rough information about gene location is sometimes available also, but these data must be used with caution because recombination is not equally likely at all places on the chromosome. Thus the genetic map, compared to the physical map, stretches in some places and compresses in others, as though it were drawn on a rubber band.

The degree of difficulty in finding a disease gene of interest depends largely on what information is already known about the gene and, especially, on what kind of DNA alterations cause the disease. Spotting the disease gene is very difficult when disease results from a single altered DNA base; sickle cell anemia is an example of such a case, as are probably most major human inherited diseases. When disease results from a large DNA rearrangement, this anomaly can usually be detected as alterations in the physical map of the region or even by direct microscopic examination of the chromosome. The location of these alterations pinpoints the site of the gene.

Identifying the gene responsible for a specific disease without a map is analogous to finding a needle in a haystack. Actually, finding the gene is even more difficult, because even close up, the gene still looks like just another piece of hay. However, maps give clues on where to look; the finer the map's resolution, the fewer pieces of hay to be tested.

Once the neighborhood of a gene of interest has been identified, several strategies can be used to find the gene itself. An ordered library of the gene neighborhood can be constructed if one is not already available. This library provides DNA fragments that can be screened for additional polymorphisms, improving the genetic map of the region and further restricting the possible gene location. In addition, DNA fragments from the region can be used as probes to search for DNA sequences that are expressed (transcribed to RNA) or conserved among individuals. Most genes will have such sequences. Then individual gene candidates must be examined. For example, a gene responsible for liver disease is likely to be expressed in the liver and less likely in other tissues or organs. This type of evidence can further limit the search. Finally, a suspected gene may need to be sequenced in both healthy and affected individuals. A consistent pattern of DNA variation when these two samples are compared will show that the gene of interest has very likely been found. The ultimate proof is to correct the suspected DNA alteration in a cell and show that the cell's behavior reverts to normal.

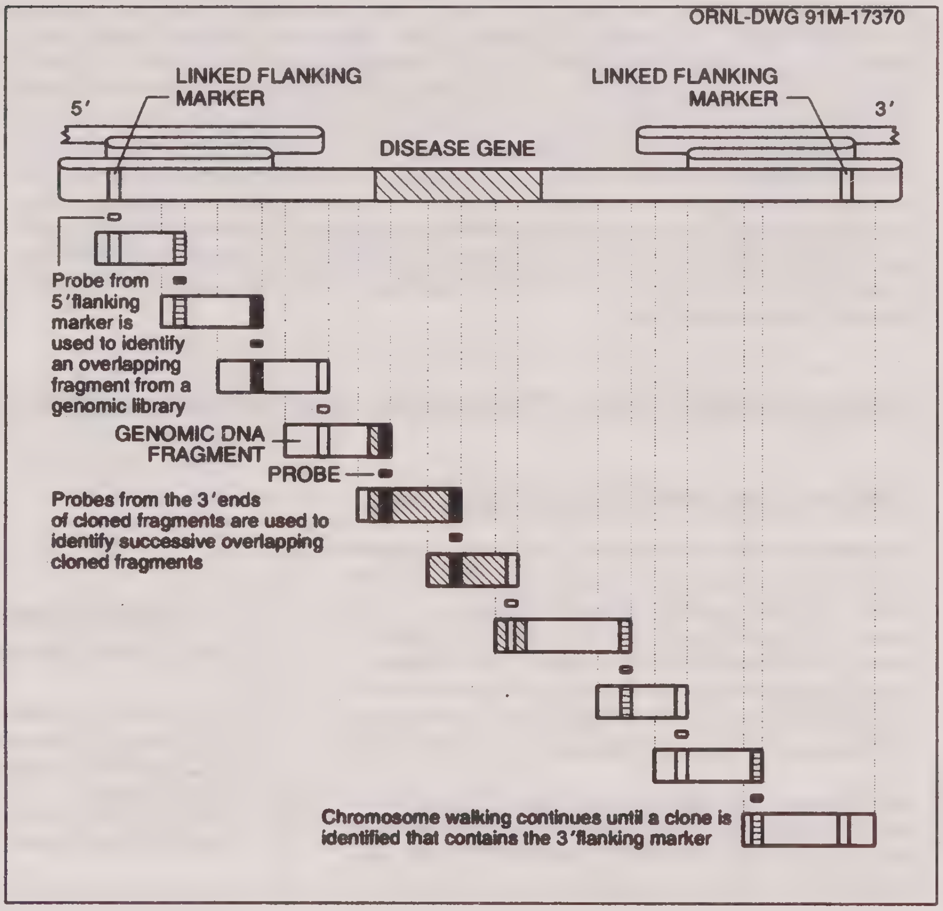

ORNL-DWG 91M-17370
LINKED FLANKING MARKER
5′
LINKED FLANKING MARKER
3′
DISEASE GENE
Probe from 5′flanking marker is used to identify an overlapping fragment from a genomic library
GENOMIC DNA FRAGMENT
PROBE
Probes from the 3′ends of cloned fragments are used to identify successive overlapping cloned fragments
Chromosome walking continues until a clone is identified that contains the 3′flanking marker

Fig. 13. Cloning a Disease Gene by Chromosome Walking. *After a marker is linked to within 1 cM of a disease gene, chromosome walking can be used to clone the disease gene itself. A probe is first constructed from a genomic fragment identified from a library as being the closest linked marker to the gene. A restriction fragment isolated from the end of the clone near the disease locus is used to reprobe the genomic library for an overlapping clone. This process is repeated several times to walk across the chromosome and reach the flanking marker on the other side of the disease-gene locus. (Source: see Fig. 11.)*

Model Organism Research

Most mapping and sequencing technologies were developed from studies of nonhuman genomes, notably those of the bacterium *Escherichia coli,* the yeast *Saccharomyces cerevisiae,* the fruit fly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans*, and the laboratory mouse *Mus musculus.* These simpler systems provide excellent models for developing and testing the procedures needed for studying the much more complex human genome.

A large amount of genetic information has already been derived from these organisms, providing valuable data for the analysis of normal gene regulation, genetic diseases, and evolutionary processes. Physical maps have been completed for *E. coli*, and extensive overlapping clone sets are available for *S. cerevisiae* and *C. elegans.* In addition, sequencing projects have been initiated by the NIH genome program for *E. coli, S. cerevisiae*, and *C. elegans.*

Mouse genome research will provide much significant comparative information because of the many biological and genetic similarities between mouse and man. Comparisons of human and mouse DNA sequences will reveal areas that have been conserved during evolution and are therefore important. An extensive database of mouse DNA sequences will allow counterparts of particular human genes to be identified in the mouse and extensively studied. Conversely, information on genes first found to be important in the mouse will lead to associated human studies. The mouse genetic map, based on morphological markers, has already led to many insights into human biology. Mouse models are being developed to explore the effects of mutations causing human diseases, including diabetes, muscular dystrophy, and several cancers. A genetic map based on DNA markers is presently being constructed, and a physical map is planned to allow direct comparison with the human physical map.

Informatics: Data Collection and Interpretation

Collecting and Storing Data

The reference map and sequence generated by genome research will be used as a primary information source for human biology and medicine far into the future. The vast amount of data produced will first need to be collected, stored, and distributed. If compiled in books, the data would fill an estimated 200 volumes the size of a Manhattan telephone book (at 1000 pages each), and reading it would require 26 years working around the clock (Fig.14).

Because handling this amount of data will require extensive use of computers, database development will be a major focus of the Human Genome Project. The present challenge is to improve database design, software for

HUMAN GENETIC DIVERSITY:
The Ultimate Human Genetic Database

- Any two individuals differ in about 3×10^6 bases (0.1%).
- The population is now about 5×10^9.
- A catalog of all sequence differences would require 15×10^{15} entries.
- This catalog may be needed to find the rarest or most complex disease genes.

database access and manipulation, and data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. Databases need to be designed that will accurately represent map information (linkage, STSs, physical location, disease loci) and sequences (genomic, cDNAs, proteins) and link them to each other and to bibliographic text databases of the scientific and medical literature.

Interpreting Data

New tools will also be needed for analyzing the data from genome maps and sequences. Recognizing where genes begin and end and identifying their exons, introns, and regulatory sequences may require extensive comparisons with sequences from related species such as the mouse to search for conserved similarities (homologies). Searching a database for a particular DNA sequence may uncover these homologous sequences in a known gene from a model organism, revealing insights into the function of the corresponding human gene.

Correlating sequence information with genetic linkage data and disease gene research will reveal the molecular basis for human variation. If a newly identified gene is found to code for a flawed protein, the altered protein must be compared with the normal version to identify the specific abnormality that causes disease. Once the error is pinpointed, researchers must try to determine how to correct it in the human body, a task that will require knowledge about how the protein functions and in which cells it is active.

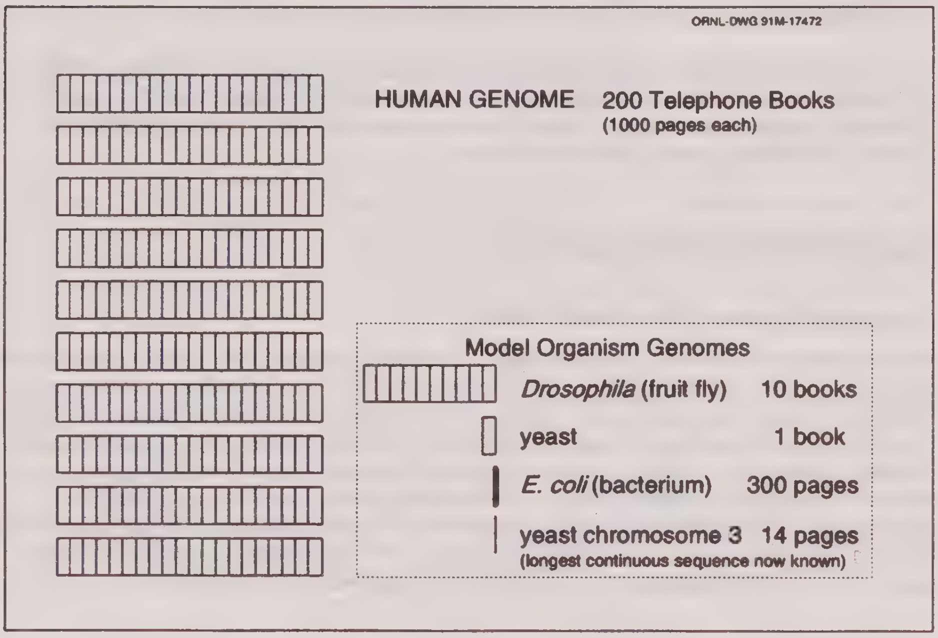

ORNL-DWG 91M-17472
HUMAN GENOME
200 Telephone Books
(1000 pages each)
Model Organism Genomes
Drosophila (fruit fly)
10 books
yeast
1 book
E. coli (bacterium)
300 pages
yeast chromosome 3
14 pages
(longest continuous sequence now known)

Fig. 14. Magnitude of Genome Data. *If the DNA sequence of the human genome were compiled in books, the equivalent of 200 volumes the size of a Manhattan telephone book (at 1000 pages each) would be needed to hold it all. New data-analysis tools will be needed for understanding the information from genome maps and sequences.*

Correct protein function depends on the three-dimensional (3D), or folded, structure the proteins assume in biological environments; thus, understanding protein structure will be essential in determining gene function. DNA sequences will be translated into amino acid sequences, and researchers will try to make inferences about functions either by com-paring protein sequences with each other or by comparing their specific 3-D structures (Fig. 15).

Because the 3-D structure patterns (motifs) that protein molecules assume are much more evolutionarily conserved than amino acid sequences, this type of homology search could prove more fruitful. Particular motifs may serve similar functions in several different proteins, information that would be valuable in genome analyses. Currently, however, only a few protein motifs can be recognized at the sequence level. Continued development of analytic capabilities to facilitate grouping protein sequences into motif families will make homology searches more successful.

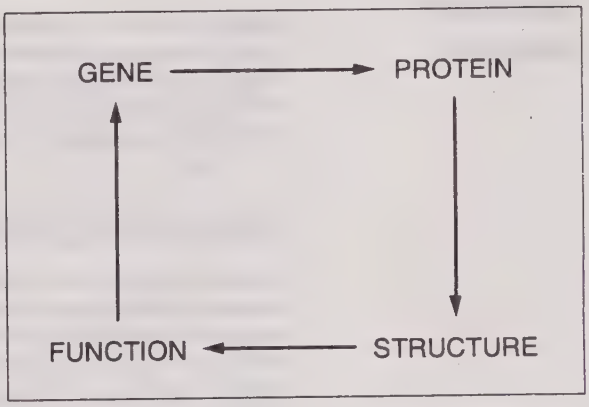


Fig. 15. Understanding Gene Function. *Understanding how genes function will require analyses of the 3-D structures of the proteins for which the genes code.*

Mapping Databases

The Genome Data Base (GDB), located at Johns Hopkins University (Baltimore, Maryland), provides location, ordering, and distance information for human genetic markers, probes, and contigs linked to known human genetic disease. GDB is presently working on incorporating physical mapping data. Also at Hopkins is the Online *Mendelian Inheritance in Man* database, a catalog of inherited human traits and diseases.

The Human and Mouse Probes and Libraries Database (located at the American *Type Culture* Collection in Rockville, Maryland) and the GBASE mouse database (located at Jackson Laboratory, Bar Harbor, Maine) include data on RFLPs, chromosomal assignments, and probes from the laboratory mouse.

Sequence Databases

Nucleic Acids (DNA and RNA)

Public databases containing the complete nucleotide sequence of the human genome and those of selected model organisms will be one of the most useful products of the Human Genome Project. Four major public databases now store nucleotide sequences: GenBank and the Genome Sequence DataBase (GSDB) in the United States, European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database in the United Kingdom, and the DNA Database of Japan (DDBJ). The databases collaborate to share sequences, which are compiled from direct author submissions and journal scans. The four databases now house a total of almost 200 Mb of sequence. Although human sequences predominate, more than 8000 species are represented. [Paragraph updated July 1994]

Proteins

The major protein sequence databases are the Protein Identification Resource (National Biomedical Research Foundation), Swissprot, and GenPept (both distributed with GenBank). In addition to sequence information, they contain information on protein motifs and other features of protein structure.

Impact of the Human Genome Project

The atlas of the human genome will revolutionize medical practice and biological research into the 21st century and beyond. All human genes will eventually be found, and accurate diagnostics will be developed for most inherited diseases. In addition, animal models for human disease research will be more easily developed, facilitating the understanding of gene function in health and disease.

Researchers have already identified single genes associated with a number of diseases, such as cystic fibrosis, Duchenne muscular dystrophy, myotonic dystrophy, neurofibromatosis, and retinoblastoma. As research progresses, investigators will also uncover the mechanisms for diseases caused by several genes or by a gene interacting with environmental factors. Genetic susceptibilities have been implicated in many major disabling and fatal diseases including heart disease, stroke, diabetes, and several kinds of cancer. The identification of these genes and their proteins will pave the way to more-effective therapies and preventive measures. Investigators determining the underlying biology of genome organization and gene regulation will also begin to understand how humans develop from single cells to adults, why this process sometimes goes awry, and what changes take place as people age.

New technologies developed for genome research will also find myriad applications in industry, as well as in projects to map (and ultimately improve) the genomes of economically important farm animals and crops.

While human genome research itself does not pose any new ethical dilemmas, the use of data arising from these studies presents challenges that need to be addressed before the data accumulate significantly. To assist in policy development, the ethics component of the Human Genome Project is funding conferences and research projects to identify and consider relevant issues, as well as activities to promote public awareness of these topics.

Glossary

Portions of the glossary text were taken directly or modified from definitions in the U.S. Congress Office of Technology Assessment document: *Mapping Our Genes—The Genome Projects: How Big, How Fast?* OTA-BA-373, Washington, D.C.: U.S. Government Printing Office, April 1988.

Adenine (A): A nitrogenous base, one member of the *base pair* A-T (adenine-*thymine*).

Alleles: Alternative forms of a genetic *locus*; a single allele for each locus is inherited separately from each parent (e.g., at a locus for eye color the allele might result in blue or brown eyes).

Amino acid: Any of a class of 20 molecules that are combined to form *proteins* in living things. The sequence of amino acids in a protein and hence protein function are determined by the *genetic code*.

Amplification: An increase in the number of copies of a specific DNA fragment; can be in vivo or in vitro. See *cloning*, *polymerase chain reaction*.

Arrayed library: Individual primary recombinant clones (hosted in *phage*, *cosmid*, *YAC*, or other *vector)* that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified by the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications, including screening for a specific *gene* or genomic region of interest as well as for *physical mapping*. Information gathered on individual clones from various genetic *linkage* and *physical map* analyses is entered into a relational database and used to construct physical and genetic *linkage maps* simultaneously; clone identifiers serve to interrelate the multilevel maps. Compare *library*, *genomic library*.

Autoradiography: A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel *electrophoresis*.

Autosome: A *chromosome* not involved in sex determination. The *diploid* human *genome* consists of 46 chromosomes, 22 pairs of autosomes, and 1 pair of *sex chromosomes* (the X and Y chromosomes).

Bacteriophage: See *phage*.

Base pair (bp): Two nitrogenous bases (*adenine* and *thymine* or *guanine* and *cytosine*) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Base sequence: The order of *nucleotide* bases in a DNA molecule.

Base sequence analysis: A method, sometimes automated, for determining the *base sequence*.

Biotechnology: A set of biological techniques developed through basic research and now applied to research and product development. In particular, the use by industry of *recombinant DNA*, cell fusion, and new bioprocessing techniques.

bp: See *base pair*.

cDNA: See *complementary DNA*.

Centimorgan (cM): A unit of measure of *recombination* frequency. One centimorgan is equal to a 1% chance that a marker at one genetic *locus* will be separated from a marker at a second locus due to *crossing over* in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million *base pairs*.

Centromere: A specialized *chromosome* region to which spindle fibers attach during cell division.

Chromosomes: The self-replicating genetic structures of cells containing the cellular DNA that bears in its *nucleotide* sequence the linear array of *genes*. In *prokaryotes*, chromosomal DNA is circular, and the entire genome is carried on one chromosome. *Eukaryotic* genomes consist of a number of chromosomes whose DNA is associated with different kinds of *proteins*.

Clone bank: See *genomic library*.

Clones: A group of cells derived from a single ancestor.

Cloning: The process of asexually producing a group of cells (clones), all genetically identical, from a single ancestor. In *recombinant DNA technology,* the use of DNA manipulation procedures to produce multiple copies of a single *gene* or segment of DNA is referred to as cloning DNA.

Cloning vector: DNA molecule originating from a *virus*, a *plasmid,* or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are *plasmids*, *cosmids*, and *yeast artificial chromosomes*; vectors are often *recombinant* molecules containing DNA sequences from several sources.

cM: See *centimorgan.*

Code: See *genetic code.*

Codon: See *genetic code.*

Complementary DNA (cDNA): DNA that is synthesized from a *messenger RNA* template; the single-stranded form is often used as a *probe* in *physical mapping.*

Complementary sequences: *Nucleic acid base sequences* that can form a double-stranded structure by matching *base pairs*; the complementary sequence to G-T-A-C is C-A-T-G.

Conserved sequence: A *base sequence* in a DNA molecule (or an *amino acid* sequence in a *protein*) that has remained essentially unchanged throughout evolution.

Contig map: A map depicting the relative order of a linked *library* of small overlapping clones representing a complete chromosomal segment.

Glossary

Contigs: Groups of *clones* representing overlapping regions of a *genome.*

Cosmid: Artificially constructed *cloning vector* containing the *cos* gene of *phage* lambda. Cosmids can be packaged in lambda phage particles for infection into *E. coli*; this permits cloning of larger DNA fragments (up to 45 kb) than can be introduced into bacterial hosts in *plasmid* vectors.

Crossing over: The breaking during *meiosis* of one maternal and one paternal *chromosome*, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of *alleles* between chromosomes. Compare *recombination.*

Cytosine (C): A *nitrogenous base*, one member of the *base pair* G-C (*guanine* and cytosine).

Deoxyribonucleotide: See *nucleotide.*

Diploid: A full set of genetic material, consisting of paired *chromosomes*—one chromosome from each parental set. Most animal cells except the *gametes* have a diploid set of chromosomes. The diploid human *genome* has 46 chromosomes. Compare *haploid.*

DNA (deoxyribonucleic acid): The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between *base pairs* of *nucleotides.* The four nucleotides in DNA contain the bases: *adenine* (A), *guanine* (G), *cytosine* (C), and *thymine* (T). In nature, *base pairs* form only between A and T and between G and C; thus the *base sequence* of each single strand can be deduced from that of its partner.

DNA probes: See *probe.*

DNA replication: The use of existing DNA as a template for the synthesis of new DNA strands. In humans and other *eukaryotes*, replication occurs in the cell *nucleus.*

DNA sequence: The relative order of *base pairs*, whether in a fragment of DNA, a *gene*, a *chromosome*, or an entire *genome*. See *base sequence analysis.*

Domain: A discrete portion of a *protein* with its own function. The combination of domains in a single protein determines its overall function.

Double helix: The shape that two linear strands of DNA assume when bonded together.

***E. coli*:** Common bacterium that has been studied intensively by geneticists because of its small genome size, normal lack of pathogenicity, and ease of growth in the laboratory.

Electrophoresis: A method of separating large molecules (such as DNA fragments or *proteins*) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Separation is based on these differences. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids.

Endonuclease: An *enzyme* that cleaves its nucleic acid substrate at internal sites in the *nucleotide* sequence.

Enzyme: A *protein* that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction.

EST: Expressed sequence tag. See *sequence tagged site.*

Eukaryote: Cell or organism with membrane-bound, structurally discrete *nucleus* and other well-developed subcellular compartments. Eukaryotes include all organisms except *viruses*, bacteria, and blue-green algae. Compare *prokaryote.* See *chromosomes.*

Evolutionarily conserved: See *conserved sequence.*

Exogenous DNA: DNA originating outside an organism.

Exons: The *protein*-coding DNA sequences of a *gene*. Compare *introns.*

Exonuclease: An *enzyme* that cleaves *nucleotides* sequentially from free ends of a linear nucleic acid substrate.

Expressed gene: See *gene expression.*

FISH (fluorescence in situ hybridization): A *physical mapping* approach that uses fluorescein tags to detect *hybridization* of *probes* with *metaphase chromosomes* and with the less-condensed *somatic interphase* chromatin.

Flow cytometry: Analysis of biological material by detection of the light-absorbing or fluorescing properties of cells or subcellular fractions (i.e., *chromosomes*) passing in a narrow stream through a laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, sort successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet.

Flow karyotyping: Use of flow cytometry to analyze and/or separate *chromosomes* on the basis of their DNA content.

Gamete: Mature male or female reproductive cell (sperm or ovum) with a *haploid* set of *chromosomes* (23 for humans).

Gene: The fundamental physical and functional unit of heredity. A *gene* is an ordered sequence of *nucleotides* located in a particular position on a particular *chromosome* that encodes a specific functional product (i.e., a *protein* or *RNA* molecule). See *gene expression.*

Gene expression: The process by which a *gene's* coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into *mRNA* and then translated into *protein* and those that are transcribed into *RNA* but not translated into protein (e.g., *transfer* and *ribosomal RNAs*).

Glossary

Gene families: Groups of closely related *genes* that make similar products.

Gene library: See *genomic library*.

Gene mapping: Determination of the relative positions of *genes* on a DNA molecule (*chromosome* or *plasmid*) and of the distance, in *linkage* units or physical units, between them.

Gene product: The biochemical material, either *RNA* or *protein*, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease-causing alleles.

Genetic code: The sequence of *nucleotides*, coded in triplets (*codons*) along the *mRNA*, that determines the sequence of *amino acids* in *protein* synthesis. The DNA sequence of a *gene* can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the *amino acid* sequence.

Genetic engineering technologies: See *recombinant DNA technologies*.

Genetic map: See *linkage map*.

Genetic material: See *genome*.

Genetics: The study of the patterns of inheritance of specific traits.

Genome: All the genetic material in the *chromosomes* of a particular organism; its size is generally given as its total number of *base pairs*.

Genome projects: Research and technology development efforts aimed at *mapping* and *sequencing* some or all of the *genome* of human beings and other organisms.

Genomic library: A collection of *clones* made from a set of randomly generated overlapping DNA fragments representing the entire *genome* of an organism. Compare *library*, *arrayed library*.

Guanine (G): A nitrogenous base, one member of the *base pair* G-C (guanine and *cytosine*).

Haploid: A single set of *chromosomes* (half the full set of genetic material), present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. Compare *diploid*.

Heterozygosity: The presence of different *alleles* at one or more *loci* on *homologous chromosomes*.

Homeobox: A short stretch of *nucleotides* whose *base sequence* is virtually identical in all the *genes* that contain it. It has been found in many organisms from fruit flies to human beings. In the fruit fly, a homeobox appears to determine when particular groups of genes are expressed during development.

Homologies: Similarities in DNA or *protein* sequences between individuals of the same species or among different species.

Homologous chromosomes: A pair of *chromosomes* containing the same linear *gene* sequences, each derived from one parent.

Human gene therapy: Insertion of normal DNA directly into cells to correct a genetic defect.

Human Genome Initiative: Collective name for several projects begun in 1986 by DOE to (1) create an ordered set of DNA segments from known chromosomal locations, (2) develop new computational methods for analyzing genetic map and DNA sequence data, and (3) develop new techniques and instruments for detecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The national effort, led by DOE and NIH, is known as the Human Genome Project.

Hybridization: The process of joining two *complementary* strands of DNA or one each of DNA and RNA to form a double-stranded molecule.

Informatics: The study of the application of computer and statistical techniques to the management of information. In *genome* projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence information, and to predict *protein* sequence and structure from DNA sequence data.

In situ hybridization: Use of a DNA or RNA probe to detect the presence of the *complementary DNA* sequence in cloned bacterial or cultured *eukaryotic* cells.

Interphase: The period in the cell cycle when DNA is replicated in the nucleus; followed by *mitosis*.

Introns: The DNA *base sequences* interrupting the *protein*-coding sequences of a *gene*; these sequences are *transcribed* into *RNA* but are cut out of the message before it is *translated* into protein. Compare *exons*.

In vitro: Outside a living organism.

Karyotype: A photomicrograph of an individual's *chromosomes* arranged in a standard format showing the number, size, and shape of each chromosome type; used in low-resolution *physical mapping* to correlate gross chromosomal abnormalities with the characteristics of specific diseases.

kb: See *kilobase*.

Kilobase (kb): Unit of length for DNA fragments equal to 1000 *nucleotides*.

Library: An unordered collection of *clones* (i.e., cloned DNA from a particular organism), whose relationship to each other can be established by *physical mapping*. Compare *genomic library, arrayed library*.

Glossary

Linkage: The proximity of two or more *markers* (e.g., *genes*, *RFLP* markers) on a *chromosome*; the closer together the markers are, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in *prokaryotes*, *mitosis* or *meiosis* in *eukaryotes*), and hence the greater the probability that they will be inherited together.

Linkage map: A map of the relative positions of genetic *loci* on a *chromosome*, determined on the basis of how often the loci are inherited together. Distance is measured in *centimorgans* (*cM*).

Localize: Determination of the original position (*locus)* of a *gene* or other *marker* on a chromosome.

Locus (pl. loci): The position on a *chromosome* of a *gene* or other chromosome *marker*, also, the DNA at that position. The use of *locus* is sometimes restricted to mean regions of DNA that are *expressed*. See *gene expression*.

Macrorestriction map: Map depicting the order of and distance between sites at which *restriction enzymes* cleave *chromosomes*.

Mapping: See *gene mapping, linkage map, physical map.*

Marker: An identifiable physical location on a *chromosome* (e.g., *restriction enzyme cutting site*, *gene*) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined. See *RFLP, restriction fragment length polymorphism.*

Mb: See *megabase.*

Megabase (Mb): Unit of length for DNA fragments equal to 1 million *nucleotides* and roughly equal to 1 *cM.*

Meiosis: The process of two consecutive cell divisions in the *diploid* progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a *haploid* set of *chromosomes*.

Messenger RNA (mRNA): RNA that serves as a template for *protein* synthesis. See *genetic code*.

Metaphase: A stage in *mitosis* or *meiosis* during which the *chromosomes* are aligned along the equatorial plane of the cell.

Mitosis: The process of nuclear division in cells that produces daughter cells that are genetically identical to each other and to the parent cell.

mRNA: See *messenger RNA.*

Multifactorial or multigenic disorders: See *polygenic disorders*.

Multiplexing: A *sequencing* approach that uses several pooled samples simultaneously, greatly increasing sequencing speed.

Mutation: Any heritable change in DNA *sequence*. Compare *polymorphism*.

Nitrogenous base: A nitrogen-containing molecule having the chemical properties of a base.

Nucleic acid: A large molecule composed of *nucleotide* subunits.

Nucleotide: A subunit of DNA or *RNA* consisting of a nitrogenous base (*adenine*, *guanine*, *thymine*, or *cytosine* in DNA; adenine, guanine, *uracil*, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of *nucleotides* are linked to form a DNA or RNA molecule. See *DNA, base pair, RNA*.

Nucleus: The cellular organelle in *eukaryotes* that contains the genetic material.

Oncogene: A *gene*, one or more forms of which is associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth.

Overlapping clones: See *genomic library*.

PCR: See *polymerase chain reaction*.

Phage: A *virus* for which the natural host is a bacterial cell.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., *restriction enzyme cutting sites*, *genes*), regardless of inheritance. Distance is measured in *base pairs*. For the human *genome*, the lowest-resolution *physical map* is the banding patterns on the 24 different *chromosomes*; the highest-resolution map would be the complete *nucleotide* sequence of the chromosomes.

Plasmid: Autonomously replicating, extrachromosomal circular DNA molecules, distinct from the normal bacterial *genome* and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as *cloning vectors*.

Polygenic disorders: Genetic disorders resulting from the combined action of *alleles* of more than one *gene* (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles; thus the hereditary patterns are usually more complex than those of single-gene disorders. Compare *single-gene disorders*.

Polymerase chain reaction (PCR): A method for amplifying a DNA *base sequence* using a heat-stable *polymerase* and two 20-base *primers*, one *complementary* to the (+)-strand at one end of the sequence to be amplified and the other complementary to the (–)-strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample.

Polymerase, DNA or RNA: *Enzymes* that catalyze the synthesis of *nucleic acids* on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides.

Polymorphism: Difference in DNA sequence among individuals. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic *linkage* analysis. Compare *mutation.*

Primer: Short preexisting polynucleotide chain to which new deoxyribonucleotides can be added by DNA *polymerase.*

Probe: Single-stranded DNA or *RNA* molecules of specific base *sequence*, labeled either radioactively or immunologically, that are used to detect the *complementary* base sequence by *hybridization.*

Prokaryote: Cell or organism lacking a membrane-bound, structurally discrete *nucleus* and other subcellular compartments. Bacteria are prokaryotes. Compare *eukaryote.* See *chromosomes.*

Promoter: A site on DNA to which *RNA polymerase* will bind and initiate *transcription.*

Protein: A large molecule composed of one or more chains of *amino acids* in a specific order; the order is determined by the *base sequence* of *nucleotides* in the *gene* coding for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs, and each protein has unique functions. Examples are hormones, *enzymes*, and antibodies.

Purine: A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine.

Pyrimidine: A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine; in RNA, cytosine and uracil.

Rare-cutter enzyme: See *restriction enzyme cutting site.*

Recombinant clones: *Clones* containing *recombinant DNA molecules.* See *recombinant DNA technologies.*

Recombinant DNA molecules: A combination of DNA molecules of different origin that are joined using *recombinant DNA technologies.*

Recombinant DNA technologies: Procedures used to join together DNA segments in a cell-free system (an environment outside a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular *chromosome.*

Recombination: The process by which progeny derive a combination of *genes* different from that of either parent. In higher organisms, this can occur by *crossing over.*

Regulatory regions or sequences: A DNA *base sequence* that controls *gene expression.*

Resolution: Degree of molecular detail on a *physical map* of DNA, ranging from low to high.

Restriction enzyme, endonuclease: A *protein* that recognizes specific, short *nucleotide sequences* and cuts DNA at those sites. Bacteria contain over 400 such *enzymes* that recognize and cut over 100 different DNA sequences. See *restriction enzyme cutting site.*

Restriction enzyme cutting site: A specific *nucleotide sequence* of DNA at which a particular *restriction enzyme* cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred *base pairs*), others much less frequently (*rare-cutter*, e.g., every 10,000 base pairs).

Restriction fragment length polymorphism (RFLP): Variation between individuals in DNA fragment sizes cut by specific *restriction enzymes*; *polymorphic sequences* that result in RFLPs are used as *markers* on both *physical maps* and genetic *linkage maps.* RFLPs are usually caused by *mutation* at a cutting site. See *marker.*

RFLP: See *restriction fragment length polymorphism.*

Ribonucleic acid (RNA): A chemical found in the *nucleus* and cytoplasm of cells; it plays an important role in *protein* synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including *messenger RNA*, *transfer RNA*, *ribosomal RNA*, and other small RNAs, each serving a different purpose.

Ribonucleotides: See *nucleotide.*

Ribosomal RNA (rRNA): A class of RNA found in the ribosomes of cells.

Ribosomes: Small cellular components composed of specialized ribosomal RNA and protein; site of protein synthesis. See *ribonucleic acid* (*RNA*).

RNA: See *ribonucleic acid.*

Sequence: See *base sequence.*

Glossary

Sequence tagged site (STS): Short (200 to 500 *base pairs*) DNA sequence that has a single occurrence in the human *genome* and whose location and base sequence are known. Detectable by *polymerase chain reaction*, STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and serve as landmarks on the developing *physical map* of the human genome. Expressed sequence tags (ESTs) are STSs derived from cDNAs.

Sequencing: Determination of the order of *nucleotides* (*base sequences*) in a DNA or *RNA* molecule or the order of *amino acids* in a *protein*.

Sex chromosomes: The X and Y *chromosomes* in human beings that determine the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a *karyotype*. Compare *autosome*.

Shotgun method: *Cloning* of DNA fragments randomly generated from a *genome*. See *library, genomic library*.

Single-gene disorder: Hereditary disorder caused by a *mutant* allele of a single *gene* (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). Compare *polygenic disorders*.

Somatic cells: Any cell in the body except *gametes* and their precursors.

Southern blotting: Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific *base sequences* by radiolabeled complementary probes.

STS: See *sequence tagged site*.

Tandem repeat sequences: Multiple copies of the same *base sequence* on a *chromosome*; used as a marker in *physical mapping*.

Technology transfer: The process of converting scientific findings from research laboratories into useful products by the commercial sector.

Telomere: The ends of *chromosomes*. These specialized structures are involved in the replication and stability of linear DNA molecules. See *DNA replication*.

Thymine (T): A nitrogenous base, one member of the *base pair* A-T (*adenine*-thymine).

Transcription: The synthesis of an *RNA* copy from a *sequence* of DNA (a *gene*); the first step in *gene expression*. Compare *translation*.

Transfer RNA (tRNA): A class of *RNA* having structures with triplet *nucleotide* sequences that are *complementary* to the triplet nucleotide coding sequences of *mRNA*. The role of tRNAs in protein synthesis is to bond with *amino acids* and transfer them to the ribosomes, where proteins are assembled according to the genetic code carried by mRNA.

Transformation: A process by which the genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its *genome.*

Translation: The process in which the genetic code carried by mRNA directs the synthesis of *proteins* from *amino acids.* Compare *transcription.*

tRNA: See *transfer RNA.*

Uracil: A nitrogenous base normally found in RNA but not DNA; uracil is capable of forming a *base pair* with *adenine.*

Vector: See *cloning vector.*

Virus: A noncellular biological entity that can reproduce only within a host cell. Viruses consist of *nucleic acid* covered by *protein*; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus.

VLSI: Very large-scale integration allowing over 100,000 transistors on a chip.

YAC: See *yeast artificial chromosome.*

Yeast artificial chromosome (YAC): A vector used to clone DNA fragments (up to 400 kb); it is constructed from the telomeric, centromeric, and replication origin sequences needed for replication in yeast cells. Compare *cloning vector, cosmid.*

Wellcome Library

ISBN 0-16-057661-X

90000

9 780160 576614